Universidad Autónoma de Madrid
Escuela Politécnica Superior

Doble Grado en Matemáticas
e Ingeniería Informática

Trabajo Fin de Grado

MODELOS BAYESIANOS PARA INTERPRETACIÓN FORENSE

Autor: Javier Santos Lorenzo
Tutor: Daniel Ramos Castro

Junio 2018

MODELOS BAYESIANOS PARA INTERPRETACIÓN FORENSE

Autor: Javier Santos Lorenzo
Tutor: Daniel Ramos Castro

Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2018

Resumen

Resumen

La estadística ha demostrado ser una herramienta valiosa para la ciencia forense. En particular, es útil para abordar el problema de comparación, consistente en decidir si dos muestras pertenecen al mismo o distinto objeto. Más concretamente, este Trabajo de Fin de Grado parte de las características físicoquímicas de muestras de vidrios y propone modelos matemáticos para ayudar en su comparación.

De las distintas aproximaciones estadísticas posibles, se han utilizado modelos bayesianos. Dichos modelos se caracterizan por establecer relaciones entre probabilidades a priori y posteriori de sucesos estocásticos. La relación entre estas probabilidades puede medirse mediante el *likelihood ratio* (LR), traduciéndose de esa manera el problema de comparación en el cálculo de dicho ratio.

Existen varias publicaciones en las que se proponen distintas aproximaciones para asignar el LR, que se diferencian en las hipótesis que establecen sobre la distribución estadística de los datos a comparar y en el grado en que la incertidumbre del problema se tiene en cuenta en el modelo. En este trabajo se han escogido dos de estos modelos para proceder a su codificación y simulación:

- Modelo de dos niveles con parámetros de máxima verosimilitud, que sólo se incorpora la incertidumbre sobre la media de las observaciones. Constituye el estado del arte actual en comparación forense.
- Modelo bayesiano completo, que incorpora la incertidumbre sobre la media y la matriz de covarianzas, además de tener en cuenta otra posible incertidumbre a través de distribuciones a priori no informativas. Este nuevo modelo es la propuesta de este Trabajo Fin de Grado.

En esta memoria se recogen las bases matemáticas en que se fundamentan estos modelos y se presentan las fórmulas utilizadas en ellos, señalando las hipótesis en que se basa cada uno.

Para su implementación se ha utilizado un entorno basado en MATLAB, herramienta de software matemático que incluye las funciones de distribución utilizadas en los modelos elegidos.

El núcleo de la memoria contiene la descripción detallada de los experimentos realizados y sus resultados, utilizando para ello tanto cuadros con datos numéricos como diferentes representaciones gráficas. Para cada uno de los modelos propuestos:

- Se ha medido el LR obtenido al comparar muestras del mismo y de distintos objetos, utilizando varias bases de datos de prueba.
- Se han utilizado curvas de entropía cruzada empírica (ECE) (véase [1]) para evaluar y representar la bondad de los valores del LR obtenido.

- Se ha variado el número de objetos y muestras por objeto para analizar su influencia en los resultados.

A la vista de los resultados obtenidos, ambos modelos han demostrado su validez para las bases de datos de prueba utilizadas, ya que la entropía empírica muestra una adecuada correspondencia entre el valor de LR calculado y las etiquetas de *"ground-truth"*.

En las conclusiones de la memoria se realizan las observaciones oportunas sobre la bondad de los resultados de cada modelo, según las características de cada experimento, y se proponen líneas de trabajo para seguir investigando sobre ellos de cara a su utilización práctica por la ciencia forense.

Palabras Clave

Bayesiano, Ciencia forense, Cllr, Inferencia distinta fuente, Inferencia misma fuente, Min_Cllr, Probabilidad a priori, Probabilidad a posteriori, Ratio de verosimilitudes (LR).

Abstract

Statistics have proved themselves as a valuable tool to support forensic science. In particular, when it comes to address the problem of comparison, consisting of deciding whether two samples belong to the same object or not. More specifically, this Final Degree Project takes into account the physicochemical characteristics of pieces of glass and proposes models that can help in their comparison.

There exist several statistical approaches. Bayesian models will be used, which provide relationships between a priori and posteriori probabilities of stochastic events, which can be measured by the likelihood ratio (LR), thus translating the comparison problem into the evaluation of this ratio.

Different equations have been proposed to calculate the likelihood ratio, differing in the underlying hypotheses about the statistical distribution of the data to be compared and in the degree that the uncertainty of the problem is taken into account in them. Two of these models have been simulated in this project:

- Between-object distribution with kernel density estimation model. This model, which constitutes the current state of the art in forensic comparison, only takes into account the uncertainty about the average of the observations.
- Fully bayesian model. It incorporates the uncertainty about the mean and the covariance. Furthermore, it takes into account other possible uncertainty through non-informative prior distributions. This new model is the proposal of this project.

This report describes the mathematical foundations these models are based on, including the equations they use, and the hypotheses about data that distinguish them.

MATLAB has been used to implement these models. It is a mathematical software tool with a wide range of libraries that include the distribution functions used in the chosen models.

The experiments conducted for both models, along with their results, are described in detail, using numeric tables and different graphic representations.

In these experiments, for the proposed models:

- LR values has been obtained comparing samples taken from the same and different objects (within-source and between-source). Different test databases have been used for this purpose.
- Empirical Cross-Entropy (ECE) curves (see [1]) have been used to evaluate and depict the adequacy of the LR values obtained in the previous experiments.
- The dependence of the results on the number of objects and samples has also been analysed.

According to the results obtained with the test databases, both models have proved their validity. In fact, the empirical entropy shows a good correspondence between the calculated LR value and the "*ground-truth*"labels.

The conclusions of the report include observations about the adequacy of the results provided by each model, according to the characteristics of each experiment.

In addition, they provide suggestions about further refinements for their practical use by forensic science.

Key words

Bayesian, Between-source variation, Cllr, Forensic Science, Likelihood ratio (LR), Min_Cllr, Posterior probability, Prior probability, Within-source variation.

Agradecimientos

En primer lugar, me gustaría agradecer a mi tutor, Daniel Ramos, tanto la posibilidad de realizar este trabajo como la atención y ayuda recibidas a lo largo de estos meses.

También tengo que agradecer a mis padres y a mi hermano su apoyo incondicional a lo largo de la carrera, especialmente en los momentos más difíciles. Gracias a ellos he aprendido que con esfuerzo, organización y trabajo se puede lograr aquello que te propongas.

Índice general

Índice de Figuras

Índice de Tablas

1 Introducción

1.1. Motivación

La **ciencia forense** permite la utilización de una serie de métodos y prácticas científicas en el contexto de un proceso legal. Una parte importante de la ciencia forense se basa en la evaluación y análisis de las características de las pruebas encontradas tanto en la escena de un crimen como en un sospechoso. Estas pruebas pueden ser de tipos muy diferentes, ya sean sustancias ilegales, huellas dactilares, sangre o diferentes fragmentos de fibras, vidrios, plásticos o residuos orgánicos.

Dos son los problemas típicos que debe resolver: el problema de comparación (determinar si dos muestras corresponden al mismo objeto) y el de clasificación. Este trabajo se centra en el primero de ellos, aplicándose, en concreto, a muestras de vidrio, con el fin de determinar si pertenecen o no a un objeto de control.

Para resolver el problema de comparación no basta con analizar sus características morfológicas (color, grosor o densidad), sino que es necesario realizar un análisis de las características fisicoquímicas de las muestras, en este caso, fragmentos de cristales. Estas características constituyen los datos que pueden ser interpretados con la ayuda de métodos estadísticos.

La estadística juega dos papeles principales en la ciencia forense:

1. El primero se plantea durante la etapa de la investigación, previa a la identificación de un sospechoso, muchas veces en la propia escena del crimen. Por ejemplo, el análisis de las muestras de ADN obtenidos a partir de los restos encontrados en el escenario del crimen (pelo, restos de piel, semen...), que permite a la policía identificar un posible culpable.

2. El segundo ocurre en la etapa del juicio, en la que la estadística puede ayudar en la evaluación de las pruebas. Por ejemplo, una vez identificado un sospechoso, el análisis de restos de tierra en las suelas de sus zapatos para confirmar que estuvo presente en el lugar del crimen. En este segundo rol es en el que se encuadra este trabajo.

Dentro de la Estadística, suele distinguirse entre lo que se denomina Estadística frecuentista y la **Estadística bayesiana**. A diferencia de la primera, la inferencia bayesiana realiza asignaciones de probabilidad sobre los parámetros; esto es posible ya que la incertidumbre debida a

la falta de conocimiento sobre los parámetros (incertidumbre epistémica) puede cuantificarse a través de la probabilidad bayesiana.

De hecho, la inferencia bayesiana describe cómo la adquisición de los datos modifica (normalmente, reduce) la incertidumbre sobre un parámetro. Para una descripción más completa véase, por ejemplo, [14].

La estadística bayesiana permite incorporar hipótesis científicas en el análisis, pudiendo aplicarse a problemas demasiado complejos para los métodos convencionales. Los métodos bayesianos permiten abordar problemas encontrados en los informes científicos y toma de decisiones judiciales, donde se debe obtener la máxima información posible a partir de los datos.

A la hora de evaluar las evidencias hay que tener en cuenta las siguientes consideraciones:

- Posibles fuentes de error o incertidumbre, incluyendo variaciones en la medida de las características dentro de los objetos recuperados y/o de control y variaciones en la medida de las características entre diferentes objetos en la población relevante.
- La correlación entre las diferentes características de un objeto, cuando se analiza más de una, así como la similitud entre el material recuperado y el de control.

Todos estos factores serán tenidos en cuenta mediante la utilización del Ratio de verosimilitudes (***likelihood ratio***)[1] **(LR)** calculado a partir de estos datos.

El LR está siendo cada vez más utilizado para evaluar el valor de unas pruebas, y se han creado métodos para obtenerlo directamente de datos tanto unidimensionales como multidimensionales, calculándolo a través de la variación existente entre las observaciones procedentes de la misma fuente (***within-source variation***) y las procedentes de diferentes fuentes (***between-source variation***).

A lo largo del trabajo, **se presentan dos modelos diferentes para calcular el LR, que se diferencian en las hipótesis utilizadas sobre la distribución de los datos**. El primero de ellos representa el actual estado del arte para la comparación forense, pero no funciona adecuadamente cuando se trabaja con muchas dimensiones y/o con pocos datos. Por ello, se va a proponer un modelo original, que se va a probar con datos simulados para comparar sus resultados con los del modelo más difundido.

1.2. Objetivos

Los objetivos principales que se persiguen en este trabajo son:

1. Derivación matemática de diferentes modelos que permitan resolver determinados problemas en el ámbito de la ciencia forense.
2. Codificación de los modelos propuestos en el punto anterior.
3. Análisis de las simulaciones realizadas con el código anterior, comparando la validez de los modelos y su robustez ante la falta de datos.
4. Estudio de las limitaciones de los modelos, detectando aquellas partes susceptibles de mejora.
5. Planteamiento de posibles ámbitos de mejora para los modelos propuestos, que permitan dotarles de una mayor robustez.

[1] Aunque existen diversas traducciones para este término se utiliza habitualmente en su expresión inglesa y así se utilizará en este trabajo.

1.3. Organización de la memoria

La memoria consta de los siguientes capítulos:

- Capítulo 1 – Introducción: Se motiva y plantea el problema y se indica la dirección a seguir en el trabajo.
- Capítulo 2 – Estado del arte de la estadística aplicada en la ciencia forense: Se describe la base matemática previa, a partir de la cual se generan los modelos descritos en el siguiente capítulo.
- Capítulo 3 – Descripción de los modelos utilizados: Se proponen y describen dos modelos bayesianos diferentes para afrontar el problema, señalando las hipótesis en que se basa cada uno de ellos.
- Capítulo 4 – Resultados de los experimentos: Se describen los diferentes experimentos realizados con cada uno de los modelos y se reflejan los resultados obtenidos, determinándose su validez y sus posibles limitaciones.
- Capítulo 5 – Conclusiones y trabajo futuro: Se recogen las conclusiones más relevantes de esta investigación y se indican las posibles vías de trabajo futuro.

Se incluyen al final del trabajo un glosario de los principales términos y acrónimos utilizados en el mismo y dos anexos:

- El primero de ellos contiene las distribuciones de probabilidad a las que se hace referencia en la descripción de los modelos matemáticos.
- El segundo anexo contiene los cuadros con los resultados numéricos de los diferentes experimentos.

2

Modelos bayesianos y ciencia forense: Estado del arte

Para poder evaluar las pruebas obtenidas en el ámbito de la ciencia forense es necesario presentar los datos de manera que puedan ser entendidos por personas no especialistas. En el contexto de esta evaluación, uno de los problemas que se debe abordar es la comparación de los datos fisicoquímicos que caracterizan estas pruebas.

Los datos obtenidos del análisis de los materiales recuperados, por ejemplo, de la ropa de un sospechoso, han de poder ser comparados con los obtenidos del material de control, denominado así porque su origen es conocido. Se pretende determinar si ese material recuperado, de origen desconocido, procede o no del mismo objeto que el material de control.

La aplicación de cualquier método de evaluación de estos datos ha de ser capaz de imitar el rol de un experto forense ante un tribunal de justicia. Este papel se basa en la evaluación de los datos de una prueba (o evidencia, E) en el contexto de las hipótesis de la acusación (H_1) y de la defensa (H_2) con el fin de asignar las probabilidades condicionadas $P(E|H_1)$ y $P(E|H_2)$.

La valoración de las evidencias (E), expresadas en forma de datos fisicoquímicos, se puede conseguir a través de la utilización del LR. En particular, la aproximación a través del *likelihood ratio* (LR) puede ser utilizada para abordar el problema de comparación.

Es decir, permite comparar datos obtenidos de dos objetos diferentes en el contexto de la comparación de hipótesis. La hipótesis de la acusación (H_1) representa la proposición que dos fragmentos proceden del mismo objeto. A su vez, la hipótesis de la defensa (H_2) representa que los dos fragmentos proceden de objetos diferentes.

Por ejemplo, en el contexto de una investigación forense, H_1 podría representar que el origen de los cristales de vidrio encontrados en la ropa del sospechoso procede de la ventana rota de la escena del crimen. A su vez, H_2 representaría que los fragmentos de vidrio encontrados en la ropa del sospechoso proceden de cualquier otra ventana o elemento de vidrio diferente a la de la escena del crimen.

La expresión general del LR es la siguiente:

$$LR = \frac{P(E|H_1)}{P(E|H_2)}$$

en el caso de información discreta, y

$$LR = \frac{f(E|H_1)}{f(E|H_2)}$$

en caso de datos continuos.

$P(.)$ representa la probabilidad y $f(.)$ la función de densidad de probabilidad.

Sin embargo, antes de explicar con más detalle esta aproximación, se presentará el teorema de Bayes, que se utilizará como base para la propuesta de modelos mediante los que abordar el problema.

2.1. Teorema de Bayes

Sea $A_1, A_2, \ldots, A_n$ un conjunto de sucesos mutuamente excluyentes y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades $P(B|A_i)$. El **teorema de Bayes** [19] establece que la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ es la **probabilidad a priori** de cada suceso A_i
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i
- $P(A_i|B)$ es la **probabilidad a posteriori** de A_i supuesto que se cumple B

De la expresión del Teorema de Bayes puede derivarse esta otra expresión:

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(H_1)}{P(H_2)} \frac{P(E|H_1)}{P(E|H_2)}$$

donde $P(H_1)$ y $P(H_2)$ son conocidas como probabilidades a priori y $P(E|H_1)$ y $P(E|H_2)$ como probabilidades a posteriori, cuya estimación en el ámbito forense se encuentra entre las competencias de las personas encargadas de resolver el caso, ya sean jueces, investigadores o policías. De este modo, podrán expresar sus opiniones a través de las hipótesis antes de que se analicen las pruebas.

La tarea de estas personas es determinar con claridad si determinados objetos proceden o no de la misma fuente. Esta tarea se realiza a través de la estimación de $P(H_1|E)$ y $P(H_2|E)$, lo cual puede lograrse teniendo en cuenta las probabilidades a priori y la información derivada por el experto forense en forma de LR.

2.2. Asignación mediante el LR

Como se ha indicado anteriormente, el valor de las evidencias (E) en forma de datos fisicoquímicos puede conseguirse a través de la aproximación del LR. Esto permite comparar datos obtenidos de dos objetos diferentes en el contexto de la comparación de hipótesis.

El método de evaluación que se utilice ha de apoyar la hipótesis correcta de una manera clara, es decir, que $LR >> 1$ **cuando** H_1 **sea la hipótesis correcta y** $LR << 1$ **cuando lo sea** H_2. Además, es deseable que su valor esté cercano a 1 cuando el LR apoye la hipótesis incorrecta (es decir, $LR > 1$ siendo H_2 correcto y $LR < 1$ cuando lo sea H_1).

Esto quiere decir que el método elegido no debería producir información engañosa que pudiera ser llevada a juicio, es decir, no deberían producirse ni **falsos positivos** ni **falsos negativos**. Un falso negativo ocurre cuando se determina que fragmentos procedentes de objetos diferentes proceden del mismo objeto. Del mismo modo, un falso positivo ocurre cuando se determina que fragmentos procedentes del mismo objeto proceden de objetos diferentes.

A la hora de evaluar las pruebas en forma de datos fisicoquímicos es necesario tener en cuenta diferentes cuestiones:

- Posibles fuentes de incertidumbre o error, tales como variaciones en las medidas del material recuperado y de control.
- Similitudes entre los datos de los objetos comparados.
- Correlación entre las variables en el caso de información multidimensional.
- La tipicalidad o rareza de los datos.

Todos estos aspectos han sido tenidas en cuenta en los modelos analizados a lo largo del trabajo.

Partiendo de una base de datos adecuada se puede obtener tanto la variabilidad interna del objeto (representada por la matriz de dispersión U) como la variabilidad entre objetos (representada por la matriz de dispersión C). En función de cuál sea la distribución de la variabilidad entre objetos, se llegará a un modelo final u otro, tal como se verá en el Capítulo 3.

Los datos obtenidos de m objetos diferentes, cada uno medido n veces (las muestras del objeto de las que se dispone) dan lugar a vectores p-dimensionales, donde p es el número de variables de interés. Ese vector puede escribirse de la siguiente manera:

$$x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$$

donde

$$i = 1, \ldots, m$$
$$j = 1, \ldots, n$$

Además, sean dos bases de datos (control y recuperada) con n_1 y n_2 como medidas y se necesita una comparación entre ellas. Para representar las estas medidas se empleará la siguiente notación:

$$y_{lj} = (y_{lj1}, \ldots, y_{ljp})^T$$

con $l = 1, 2$ (para las muestras de control y recuperadas respectivamente) y $j = 1, \ldots, n_l$, siendo n_l el número de medidas existentes para los objetos de control y recuperados.

En el caso de información multivariada, el LR puede calcularse utilizando la siguiente expresión:

$$LR = \frac{P(\bar{y}_1, \bar{y}_2|H_1, D)}{P(\bar{y}_1, \bar{y}_2|H_2, D)}$$

donde

- $\bar{y}_1$ e $\bar{y}_2$ representan las muestras de los objetos que se están analizando.
- H_1 representa la hipótesis de que ambas muestras proceden del mismo objeto.
- H_2 representa la hipótesis de que las muestras proceden de objetos distintos.
- D hace referencia a los datos de control.

Dado que en la expresión anterior $\bar{y}_1$ e $\bar{y}_2$ se supone que proceden de objetos diferentes, por lo que pueden considerarse independientes, el denominador puede reescribirse dando lugar a esta nueva expresión:

$$\begin{aligned} LR &= \frac{P(\bar{y}_1, \bar{y}_2|H_1, D)}{P(\bar{y}_1|H_2, D)P(\bar{y}_2|H_2, D)} \\ &= \frac{\int P(\bar{y}_1, \bar{y}_2|\theta)P(\theta|D)d\theta}{\int P(\bar{y}_1|\theta)P(\theta|D)d\theta \int P(\bar{y}_2|\theta)P(\theta|D)d\theta} \end{aligned} \tag{2.1}$$

Esta es la **expresión general que se utilizará en este trabajo para el cálculo del LR** y, de hecho, se harán referencias a ella en el siguiente capítulo. **Dependiendo de las suposiciones que se utilicen** para el cálculo de la probabilidad P **se obtendrán diferentes modelos**, de los que se presentarán dos en el capítulo 3.

2.3. LR máximo y mínimo teóricos

En algunos de los experimentos realizados en este trabajo se utilizan los **valores teóricos máximo y mínimo** del LR, que se calculan como se describe a continuación. Se parte de la siguiente ecuación:

$$P(H_1|E) = \frac{LR\left(\frac{P(H_1)}{P(H_2)}\right)}{1 + LR\left(\frac{P(H_1)}{P(H_2)}\right)} \tag{2.2}$$

En ella, $P(H_1|E)$ se sustituye por

$$P_{max} = \frac{N+1}{N+2}$$

para obtener el LR máximo y por

$$P_{min} = \frac{1}{N+2}$$

para obtener el mínimo. La justificación de estos valores está basada en la regla de sucesión de Laplace, que se enuncia seguidamente[1].

En un experimento, del que sabe a priori que puede dar como resultado un éxito o un fracaso, tras repetirlo n veces independientes se obtienen s éxitos. Se pretende determinar es la probabilidad que la próxima repetición sea nuevamente un éxito.

Para enunciarlo formalmente, sean $X_1, ..., X_{n+1}$ variables aleatorias, uniformemente distribuidas en el intervalo $[0, 1]$, condicionalmente independientes, con una **distribución de Bernoulli** con valor esperado p (es decir cada una posee una probabilidad p de ser igual a 1 y una probabilidad $1 - p$ de ser igual a cero 0). Bajo estas hipótesis:

$$P(X_{n+1} = 1 | X_1 + \ldots + X_n = s) = \frac{s+1}{n+2}$$

Además, en la fórmula (2.2) $P(H_1)$ se aproxima por

$$\frac{N_1 + 1}{N + 2}$$

que representa el porcentaje de LRs de la misma fuente, siendo N_1 el número de LRs de la misma fuente.

Por su parte, $P(H_2)$ se aproxima por

$$\frac{N_2 + 1}{N + 2}$$

que representa el porcentaje de LR_s de fuentes diferentes, siendo N_2 el número de LR_s de fuentes diferentes.

Esta asignación de probabilidad se denomina en muchos textos *probabilidad a priori empírica.*

A partir de las probabilidades a posteriori máxima y mínima, junto con las probabilidades a priori empíricas, despejando en (2.2) se obtienen el LR máximo y mínimo teóricos, cuyas valores son los siguientes:

$$LR_{max} = \frac{P_{max} \frac{P(H_1)}{P(H_2)}}{\frac{P(H_1)}{P(H_2)} - \frac{P(H_1)}{P(H_2)} P_{max}} = \frac{P_{max}}{1 - P_{max}}$$

$$LR_{min} = \frac{P_{min} \frac{P(H_1)}{P(H_2)}}{\frac{P(H_1)}{P(H_2)} - \frac{P(H_1)}{P(H_2)} P_{min}} = \frac{P_{min}}{1 - P_{min}}$$

[1] Una explicación más detallada puede encontrarse en [8].

3

Descripción de los modelos utilizados

3.1. Modelo de dos niveles con parámetros de máxima verosimilitud (Aitken-Lucy)

El primer modelo utilizado en este trabajo para el cálculo del LR ha sido desarrollado tanto en [2] como en [21][1]. En ambos textos se parte de desarrollos previos en los que se establecía la hipótesis de que los datos se distribuyen normalmente, que en este trabajo es sustituida por la hipótesis de que los datos siguen una distribucion *kernel*.

Inicialmente, se parte de una población de p características de elementos de un material; en este proyecto, los elementos serán fragmentos de vidrio y sus características serán medidas de su composición química, en forma de ratios entre concentraciones de elementos químicos en la muestra (por ejemplo, $log(Ca/Fe)$).

En la descripción que sigue se han integrado las notaciones utilizadas en ambos textos, en aras de una mejor comprensión.

Se dispone de n medidas ($n \geq 2$) de estas características para una muestra aleatoria de m objetos de esa población. Estos datos se representan como

$$x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T,$$

con

$$i = 1, \ldots, m$$
$$j = 1, \ldots, n$$

La media de cada objeto será:

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$$

[1]En lo sucesivo, en aras de la simplicidad, tanto en el texto como en los pies de cuadros y figuras se hablará de primer modelo o modelo de Aitken-Lucy (en atención a los autores del primer texto arriba mencionado).

Por su parte, la información recuperada y de control serán:

$$\{y_l\} = y_{lj},$$

Con

$$j = 1, \ldots, n_l$$
$$l = 1, 2$$

Donde

$$y_{lj} = (y_{lj1}, \ldots, y_{ljp})^T$$

Siendo $\bar{y}_{l,k}$ la media de cada una de las variables, con

$$k = 1, \ldots, p$$

La media de cada objeto será:

$$\bar{y}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} y_{lj}$$

Este modelo contempla dos orígenes de variabilidad, que se referencian a continuación, estableciendo al tiempo las hipótesis que lo caracterizan:

1. Variabilidad entre fragmentos del mismo objeto (*within-source variation*), que **se supone constante y distribuida normalmente**. Su matriz de covarianzas se denominará U.
2. Variabilidad entre objetos (*between-source variation*), que en este modelo **se supondrá sigue una distribución *kernel***. Su matriz de covarianzas se denominará C.

Dado un conjunto de datos, que en este caso serán los vectores de media de la población de partida $(\bar{x}_1, \ldots, \bar{x}_m)$, la función de densidad *kernel* (*Kernel* ***PDF***) será una función de densidad normal multidimensional con medias $\bar{x}_i$ y matriz de covarianza h^2C, donde

$$h = h_{opt} = \left(\frac{4}{m(2p+1)} \right)^{\frac{1}{p+4}}$$

Bajo estos supuestos, la expresión del LR establecida en la ecuación (2.1) tendría como numerador:

$$f_0(\bar{y}_1, \bar{y}_2 | U, C) = (2\pi)^{-p} |D_1|^{-1/2} |D_2|^{-1/2} |C|^{-1/2} (mh^p)^{-1} |D_1^{-1} + D_2^{-1} + (h^2C)^{-1}|^{-1/2}$$
$$\times exp\left\{ -\frac{1}{2} (\bar{y}_1 - \bar{y}_2)^T (D_1 + D_2)^{-1} (\bar{y}_1 - \bar{y}_2) \right\}$$
$$\times \sum_{i=1}^{m} exp\left[-\frac{1}{2} (y^* - \bar{x}_i)^T \left\{ (D_1^{-1} + D_2^{-1})^{-1} (h^2C) \right\}^{-1} (y^* - \bar{x}_i) \right]$$

Su denominador sería:

$$f_1(\bar{y}_1, \bar{y}_2|U, C) = (2\pi)^{-p}|C|^{-1}(mh^p)^{-2}\prod_{l=1}^{2}[|D_l|^{-1/2}|D_l^{-1} + (h^2C)^{-1}|^{-1/2}$$
$$\times \sum_{i=1}^{m} exp\{-\frac{1}{2}(\bar{y}_l - \bar{x}_i)^T(D_l + h^2C)^{-1}(\bar{y}_l - \bar{x}_i)\}]$$

Donde

$$y^* = (D_1^{-1} + D_2^{-1})^{-1}(D_1^{-1}\bar{y}_1 + D_2^{-1}\bar{y}_2)$$

$$D_l = \frac{U_l}{n_l}$$

La variabilidad interna U se estima mediante la expresión:

$$\tilde{U} = \frac{S_w}{m(n-1)}$$

Por su parte, la variabilidad entre objetos C se estima mediante la expresión:

$$\tilde{C} = \frac{S^*}{m-1} - \frac{S_w}{mn(n-1)}$$

siendo S_w la suma de cuadrados de las desviaciones de cada medida x_{ij} con relación a la media de cada uno de los m objetos $\bar{x}_i$:

$$S_w = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$$

A su vez, S^* es la suma de cuadrados de las desviaciones de la media de cada objeto con respecto a la media global $\bar{x}$, reflejadas en las siguientes expresiones:

$$S^* = \sum_{i=1}^{m}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

$$\bar{x} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}$$

3.2. Modelo bayesiano completo (NFB)

La base teórica de este segundo modelo está descrita en detalle en [10][2]. Al igual que en el modelo anterior, se va a partir de la fórmula del LR utilizada en (2.1):

$$LR = \frac{\int P(\bar{x}, \bar{y}|\theta)P(\theta|D)d\theta}{\int P(\bar{x}|\theta)P(\theta|D)d\theta \int P(\bar{y}|\theta)P(\theta|D)d\theta}$$

[2]En lo sucesivo, se hablará de segundo modelo o modelo NFB.

Como se verá a continuación, se establecerán hipótesis diferentes a las planteadas en el modelo anterior, que conducirán a una aproximación distinta.

Se dispone de N muestras independientes

$$X = [x_1, \ldots, x_N]$$

parametrizadas por

$$\theta = (m, V),$$

donde m el vector de medias y V es la matriz de covarianzas.

$$p(x_1, \ldots, x_N) = p(x_1|\theta), \ldots, p(x_N|\theta)p(\theta)$$

En este modelo se establecen dos hipótesis:

1. La primera es que $p(X|\theta)$ **sigue una distribución gaussiana**, es decir:

$$p(X|m, V) = N(x; m, V) = \frac{1}{|2\pi V|^{1/2}} exp\left(-\frac{1}{2}(x-m)^T V^{-1}(x-m)\right)$$

2. La segunda hipótesis es que $p(\theta|D)$ **sigue una distribución *gaussian gamma***, y por tanto puede expresarse como:

$$p(m, V) = p(m|V)p(V)$$
$$= \alpha|2\pi V|^{-1/2}|V|^{-(d+1)/2}$$

Como se explica en [10], la resolución del producto de las integrales que figuran en el denominador de la expresión del LR viene dada por $p(x|X)$ (denominada distribución posterior predictiva):

$$p(x|X) = \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left(\frac{N}{N+1}\right)^{d/2} \frac{|\pi S|^{N/2}}{|\pi S'|^{(N+1)/2}}$$

$$= \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left|\frac{NS^{-1}}{\pi(N+1)}\right|^{1/2} \left(\frac{N(x-\bar{x})^T S^{-1}(x-\bar{x})}{N+1} + 1\right)^{-(N+1)/2}$$

En esa ecuación

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$S = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$
$$S' = S + \frac{N}{N+1}(x-\bar{x})(x-\bar{x})^T$$

Lo que permite aproximarlo como

$$p(x|X) \sim \tau(\bar{x}, \frac{N+1}{N}S, N+1) \qquad if \;\; N > d \tag{3.1}$$

Siendo d la longitud del vector.

Por tanto, el denominador de la ecuación que representa el LR sigue una distribución *T de student multivariada* τ (véase su descripción en el Anexo 1).

Cuando el número de muestras N es muy grande ($N \to \infty$), esta distribución puede aproximarse por una gaussiana de media $\bar{x}$ y varianza S/N. Esta misma hipótesis se aplicará, de manera análoga en el numerador.

Por su parte, el numerador de la ecuación del LR también sigue una distribución *T de student multivariada*

$$\tau(\mu, \frac{N-m}{N-m-1}\tilde{S}, N-m) \tag{3.2}$$

donde

- N es el número total de muestras,
- m es el número de objetos,
- μ es el vector de medias

  $$\mu = \left| \begin{array}{c} \bar{x} \\ \bar{x} \end{array} \right|$$

  siendo $\bar{x}$ la media global de todos los objetos,
- $\tilde{S}$ es la matriz

  $$\tilde{S} = \left| \begin{array}{cc} S & S - S_w \\ S - S_w & S \end{array} \right|$$

  siendo

  $$S_w = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T,$$

- n es el número de muestras por objeto.

La hipótesis de la que se parte al plantear este modelo es que sus resultados van a mejorar los del de Aitken-Lucy, dado que, a diferencia de éste, incorpora la incertidumbre de los datos tanto en las medias como en las covarianzas.

Por otra parte, hay que tener en cuenta que asume normalidad, por lo que es previsible que funcione bien en entornos simulados con dicha distribución, pero no necesariamente en casos reales.

4

Resultados de los experimentos

4.1. Entorno de desarrollo

Según se señaló en la introducción, uno de los objetivos de este trabajo es la codificación de los modelos presentados en el capítulo 3, permitiendo su desarrollo y simulación y el posterior análisis de los resultados obtenidos.

Antes de proceder al desarrollo en sí, una primera decisión a tomar es el entorno o lenguaje de programación utilizado para realizarlo. Se ha elegido MATLAB [5], una herramienta de software matemático que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio.

Pueden apuntarse varias razones para realizar esta elección:

1. Se trata de una herramienta robusta.
2. Su uso está muy difundido para desarrollos matemáticos.
3. La Universidad Autónoma dispone de licencias *campus* a disposición de estudiantes e investigadores, muy convenientes para desarrollar un trabajo fin de grado.

MATLAB tiene numerosas funcionalidades, incluso superiores a las necesarias para este proyecto. Entre sus prestaciones básicas se hallan:

- La manipulación de matrices.
- La representación de datos y funciones.
- La implementación de algoritmos.
- La creación de interfaces de usuario (GUI).
- La comunicación con programas en otros lenguajes y con otros dispositivos hardware.

Como complemento a MATLAB, para la obtención de los diagramas de dispersión que se presentan más adelante en este capítulo, se ha utilizado R (véase [15]), que es un entorno y lenguaje de programación con un enfoque al análisis estadístico.

R es una implementación de software libre del lenguaje S, pero con soporte de alcance estadístico.

4.2. Base de datos original

La realización de los experimentos requiere partir de unos datos con los que realizar los cálculos. Se ha partido de la base de datos obtenida en www.wiley.com/go/physicochemical, con datos sobre cristales, que contiene los siguientes campos:

- **Si:** concentración en silicio de la muestra de vidrio.
- **K:** concentración en potasio de la muestra.
- **Fe:** concentración en hierro de la muestra.
- **Ca:** concentración en calcio de la muestra.
- **Group:** etiqueta sobre el tipo de vidrio (edificios, parabrisas. . .). Esta variable no se tendrá en cuenta en el análisis.
- **logCaK:** logaritmo del ratio entre las concentraciones de Ca y K.
- **logCaSi:** logaritmo del ratio entre las concentraciones de Ca y Si.
- **logCaFe:** logaritmo del ratio entre las concentraciones de Ca y Fe.
- **Window:** determina las fuentes de las medidas, es decir, diferentes valores de esta variable implicarán diferentes objetos. En la base de datos se dispone de 62 objetos diferentes con 5 medidas de cada uno.

A la hora de realizar los diferentes experimentos sólo serán objeto de estudio los logaritmos de las relaciones entre las concentraciones en lugar de utilizar éstas directamente. Esto presenta dos ventajas.

1. Por una parte, permite reducir la dimensión de las variables a utilizar, pues se trabaja con ratios y no directamente con las variables.
2. Por otra, normaliza las variables, permitiendo un mejor comportamiento estadístico.

4.3. Experimentos realizados con el modelo Aitken-Lucy

4.3.1. Medida del LR en comparaciones con la misma y distinta fuente

El primer experimento consistió en realizar comparaciones, de dos tipos diferentes:

1. Por una parte, entre elementos de un mismo objeto o fuente.
2. Por otra, entre elementos de fuentes diferentes.

En ambos casos, el fin es comprobar si efectivamente los valores obtenidos para el LR corresponden a los esperados para objetos de la misma o de distinta fuente, respectivamente.

Para las **comparaciones de la misma fuente**, se compararon bloques de tres medidas de un objeto con el bloque formado por las dos muestras restantes. Por tanto, se realizaron 10 comparaciones por objeto y 620 comparaciones en total.

Para este tipo de comparaciones, la teoría establece que debería obtenerse un valor de LR $>$ 1. Como en este caso se sabe con certeza que las muestras que se comparan pertenecen al mismo objeto, los resultados de todas las comparaciones deberían responder a ese patrón y el que ocurra o no así permitirá una buena evaluación de lo adecuado que es el modelo. En la medida en que éste se basa en un método estadístico y empírico los resultados obtenidos no tienen por qué ser óptimos.

Los resultados obtenidos en la práctica se resumen en el cuadro 4.1.

Valor del LR	**% comparaciones**
< 1	0,81 %
~ 1	0,16 %
1,9 - 10	1,61 %
10 - 100	6,45 %
100 - 1.000	39,35 %
> 1.000	51,61 %

Cuadro 4.1: Modelo Aitken-Lucy: LR para comparación de la misma fuente

Como puede apreciarse en dicho cuadro, **más del 99 % de las comparaciones generan un LR > 1**, como sería de esperar (la proporción de falsos positivos es inferior al 1 %). Además, más del 90 % de los casos muestran evidencias claras para determinar que las muestras comparadas proceden de un mismo objeto.

Para las **comparaciones de fuentes diferentes** se compararon cada una de las cinco medidas de un objeto con las cinco medidas del resto de objetos, realizando en total 1891($= 62 * 61/2$) comparaciones.

Para este tipo de comparaciones, según se señaló en capítulos anteriores, se esperaría obtener un valor de LR < 1. De forma análoga a lo establecido en las comparaciones dentro del mismo objeto, pueden producirse diferencias entre los resultados esperados y los obtenidos, y esa diferencia es una medida de la adecuación del modelo.

En la práctica, los valores obtenidos para el LR fueron los reflejados en el cuadro 4.2.

En este caso, **algo más de un 6 % ofrecen un resultado diferente al esperado, es decir, generan un LR > 1**, cuando al comparar muestras procedentes de objetos diferentes el LR obtenido debería ser < 1. Se trata, por tanto, de falsos negativos. Por otro lado, más del 93 % de las comparaciones permitirían establecer con claridad que las muestras comparadas proceden de objetos diferentes.

Como puede verse, el modelo **ofrece un resultado que se ajusta a lo previsto** en un elevado porcentaje de casos. Sin embargo, **parece preciso realizar algunos ajustes en el modelo**, ya que en la ciencia forense no es conveniente una tasa de errores tan elevada.

Valor del LR	% comparaciones
»1	6,16 %
∼ 1	0,10 %
0,1 - 0,85	0,39 %
0,01 - 0,1	0,29 %
< 0,01	93,05 %

Cuadro 4.2: Modelo Aitken-Lucy: LR para comparación de distinta fuente

4.3.2. Estimación del rendimiento mediante Cllr y Min_Cllr

El experimento anterior mide la bondad del LR obtenido en general para la base de datos con la que se trabaja.

Se ha considerado interesante estimar cómo puede afectar a los resultados ofrecidos por este modelo la variación de la cantidad de información que se dispone. Para ello, es necesario realizar pruebas en las que varíe tanto el número de objetos como el de muestras por objeto.

Antes de exponer sus resultados de este nuevo experimento hay que introducir nuevos conceptos. Los valores calculados para el LR pueden representarse mediante **curvas ECE** (representación gráfica de la **entropía empírica cruzada**), que miden la precisión de un conjunto de LRs en términos de una pérdida media de información [1].

La curva ECE asociada a los valores obtenidos para el LR en el experimento recién descrito es la siguiente:

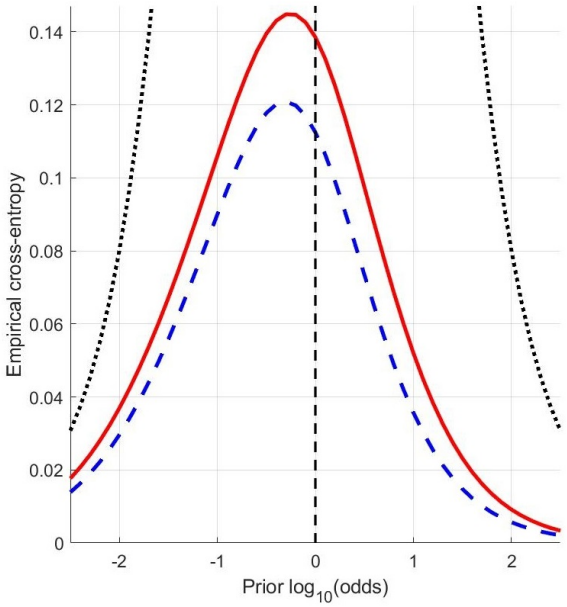

Figura 4.1: Modelo Aitken-Lucy: Curvas ECE de la base de datos original

en la que:

[1] Su complejidad hace que queden fuera del ámbito de este trabajo; se describen en profundidad en [1].

- La curva superior, dibujada en color rojo, representa a la curva empírica obtenida a partir de los datos del experimento.
- La curva azul representa la curva óptima.
- La curva discontinua negra representa los puntos con LR = 1. Por esto motivo, aquellas zonas en las que la curva roja está por encima de la discontinua reflejan que utilizar el modelo es peor que no utilizarlo.

Expresándolo de una forma más o menos simple, **la adecuación de los valores obtenidos para el LR utilizando diferentes métodos es tanto mayor cuanto más se aproxime la curva empírica a la curva óptima**.

En los siguientes experimentos se va a utilizar otro tipo de representación, en la que se compararán los valores de:

- **Cllr**: representa el valor de la curva empírica cuando corta al eje vertical. Puede considerarse una medida estándar de la precisión de un sistema que genera LRs[2].
- **Min_Cllr**: representa el valor de la curva óptima cuando corta al eje vertical. Es el valor mínimo del Cllr.

De alguna manera, **se resume la bondad de la curva ECE en esos dos valores, siendo mayor cuanto más próximos sean entre sí**, es decir, cuanto más se aproxime la curva obtenida empíricamente a la óptima. La justificación de esta afirmación puede encontrarse en [12].

El objetivo de esta nueva serie de experimentos no es sólo valorar la bondad del LR, sino determinar la influencia que tiene el número de objetos y el número de muestras en ese valor.

Por ese motivo, en cada experimento se van a realizar diferentes medidas (variando el número de objetos y el número de muestras por objeto), y en las gráficas que reflejan sus resultados se representarán los distintos valores obtenidos mediante dos planos, uno de los cuales recoge los diferentes valores del Cllr y el otro los valores del Min_Cllr.

Se trata de una representación tridimensional, en la que el eje z representa los valores medidos para cada una de esas magnitudes (Cllr y Min_Cllr), y los ejes x e y representan el número de muestras y objetos, respectivamente.

4.3.2.1. Base de datos ampliada

Para realizar estos nuevos experimentos, surge un problema adicional: la base de datos utilizada en el primer experimento tiene un número muy reducido de objetos y muestras. Esto permitiría un análisis muy limitado de la influencia de su número en el resultado, por lo que es preciso trabajar con una base de datos mayor.

Por ello, partiendo de la base de datos original, se decidió generar una nueva base de datos de mayor tamaño (aumentando el número de objetos y mediciones por objeto) para hacer con ella los siguientes experimentos. En los casos en que se requiera un número limitado de objetos y muestras, se partirá de la base de datos ampliada y se seleccionarán el número de objetos y muestras por objeto que se desee.

[2]El Cllr no utiliza directamente el valor del LR, sino su logaritmo, de ahí su denominación: *Log-Likelihood Ratio Cost*.

Obsérvese que la creación de esta base de datos ampliada puede suponer una nueva fuente de error, en la medida en la que ya no se está trabajando con datos reales sino con una simulación de los mismos, cuya distribución podría afectar a las conclusiones de los experimentos.

Para la creación de la base de datos ampliada se tendrá en cuenta tanto la variabilidad entre objetos (*between-source variation*) como la variabilidad dentro del objeto (*within-source variation*) de la base de datos original, generando aleatoriamente datos distribuidos normalmente que las tengan en cuenta. Esta base de datos tendrá en cuenta los log-ratios que se mencionaron anteriormente.

En una primera aproximación, se amplió la base de datos de 62 a 100 objetos, aumentando también el número de muestras por objeto de 5 a 9.

Con esa nueva base de datos se hizo una prueba, representando sus resultados mediante el diagrama tridimensional citado, en el que se comparan los valores de Cllr y Min_Cllr. La figura 4.2 recoge los resultados obtenidos.

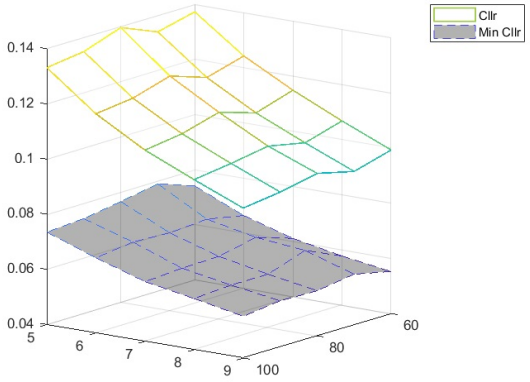

Figura 4.2: Resultados con base de datos de 100 objetos (sin limitación)

En esta gráfica se visualizan los valores obtenidos por aplicación directa del primer modelo descrito en el capítulo anterior.

Se van a comparar sus resultados con los obtenidos al imponer una limitación basada en los valores teóricos máximo y mínimo del LR descritos al final del capítulo 2. Cualquier valor del LR que supere el valor teórico máximo allí establecido será fijado a dicho valor; análogamente, los valores del LR que queden por debajo del valor teórico mínimo tomarán este valor. Se hablará, por ello, de **resultados sin y con limitación**, respectivamente. La figura 4.3 refleja esta segunda posibilidad.

Aunque no hay grandes diferencias entre los resultados obtenidos en los experimentos con y sin limitación, sí se aprecia que las diferencias entre ambos planos son ligeramente menores al aplicar la limitación, y que los valores obtenidos para Cllr y Min_Cllr son ligeramente menores, por lo que se estimó conveniente comparar ambas posibilidades en los sucesivos experimentos.

Ambas gráficas reflejan que al aumentar la cantidad de información disponible ambos planos se aproximan, es decir, conforme a lo esperado, los resultados son mejores según aumenta el número de objetos y muestras. Sin embargo, también permite observar que en el punto de mayor proximidad todavía existe una diferencia notable entre la pendiente de los planos, lo que permite suponer que estos seguirían aproximándose si se aumentara el número de muestras y, sobre todo, de objetos.

Análogamente, parece interesante ver la diferencia entre ambos planos en el extremo opuesto, cuando se reduce el número de objetos y muestras por objeto, con el fin de estimar el funciona-

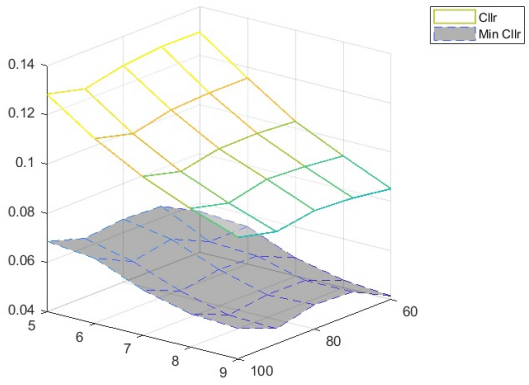

Figura 4.3: Resultados con base de datos de 100 objetos (con limitación)

miento del modelo en situaciones de poca información.

Por ello, se decidió generar una nueva base de datos, con 200 objetos y 10 muestras por objeto, a la que se denominó **base de datos ampliada**, que es la que finalmente se utilizó en la siguiente serie de experimentos.

Las figuras 4.4 y 4.5 (los **diagramas de dispersión** de las bases de datos original y ampliada, respectivamente) muestran cómo se distribuyen los datos para cada uno de los log-ratios y sus coeficientes de correlación. Obsérvese que en la segunda base de datos, estos están distribuidos normalmente, tal como se había propuesto, a diferencia de la primera. Esto provocará que el modelo NFB funcione considerablemente mejor con la base de datos simulada que con la real.

No se puede perder de vista, además, que se trata de una generación aleatoria de valores, por lo que cada ejecución puede dar lugar a unos resultados diferentes (como ocurrirá, en general, en los experimentos).

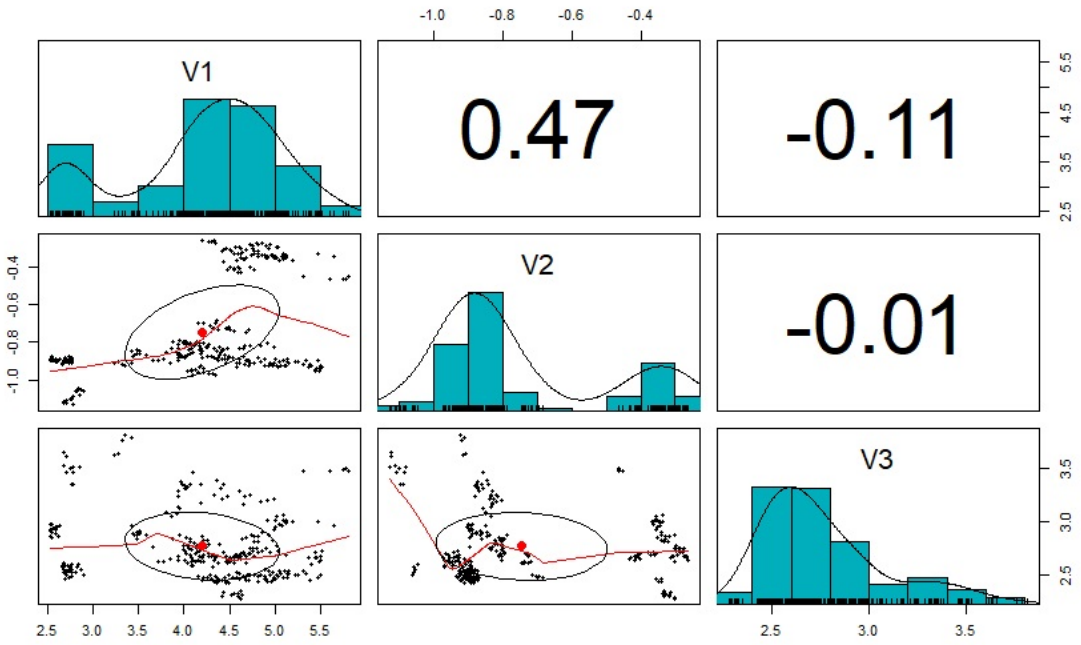

Figura 4.4: Diagrama de dispersión de la base de datos original

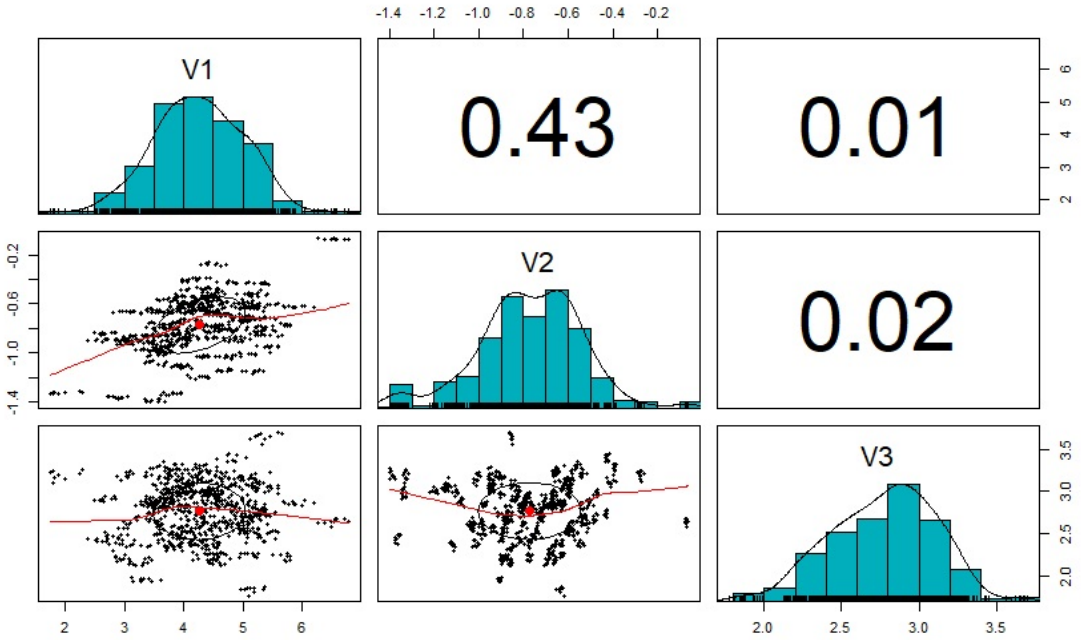

Figura 4.5: Diagrama de dispersión de la base de datos ampliada

Antes de proceder a realizar estos experimentos, se calculó el LR para la base de datos ampliada, de una manera similar a la realizada para la base de datos original en el apartado anterior. Los cuadros 4.3 y 4.4 reflejan la medida del LR para comparaciones dentro del mismo objeto y entre objetos diferentes, ambos para una generación particular de la base de datos ampliada.

Valor del LR	**% comparaciones**
< 1	7,31 %
~ 1	2,34 %
1,9 - 10	8,35 %
10 - 100	13,26 %
100 - 1.000	18,82 %
> 1.000	49,92 %

Cuadro 4.3: Base de datos ampliada: LR para comparación de la misma fuente

Si se comparan los resultados mostrados en estos cuadros con los obtenidos en el primer experimento, se observa que se obtienen distribuciones más o menos parecidas y se mantiene la conclusión allí establecida (el resultado se ajusta a lo previsto en un elevado porcentaje de casos, pero habría que estudiar cómo ajustarlo más).

También cabe insistir en que estos cuadros reflejan **los resultados de una base de datos ampliada concreta, derivada de una única ejecución del proceso de simulación**. Otras simulaciones ofrecerían resultados diferentes, aunque previsiblemente similares.

Valor del LR	**% comparaciones**
»1	2,13 %
~ 1	0,38 %
0,1 - 0,85	0,59 %
0,01 - 0,1	0,88 %
$< 0{,}01$	96,02 %

Cuadro 4.4: Base de datos ampliada: LR para comparación de distinta fuente

4.3.2.2. Resultados en función del número de objetos y muestras

Tal como se ha indicado, los resultados de las sucesivas pruebas se representarán mediante gráficas en las que se visualizan los planos de los valores de Cllr y Min_Cllr en función del número de objetos y del número de muestras por objeto (a mayor cercanía entre los planos, mejores son los resultados). Los resultados numéricos obtenidos, a partir de los cuales se construyen esos planos, están recogidos en el Anexo 2.

En la batería de experimentos cuyos resultados se presentan en este apartado se utilizó la base de datos ampliada, con 200 objetos y 10 muestras por objeto.

Se obtuvieron diferentes resultados variando:

- El número de objetos utilizado, entre 20 y 200.
- El número de muestras por objeto, con un máximo de 10.
- El número de muestras elegidas para comparar. En la primera prueba se cogerá una muestra para cada objeto y se comparará individualmente con muestras del mismo o diferente objeto (comparación de 1 en 1). A continuación, se cogerá un bloque de 2 muestras del mismo objeto y se compararán con bloques de dos muestras de un objeto, que puede ser el mismo o diferente que el de las muestras a comparar (comparación de 2 en 2).

  De la misma manera, se harán comparaciones de 3 en 3 y de 4 en 4. Según el número de muestras que compongan el bloque que se compara, variará el número mínimo de muestras que pueden utilizarse.
- Si se trata de una prueba con o sin limitación, en los términos descritos más arriba.

Dado que la base de datos con la que se trabaja está generada aleatoriamente, se ha repetido varias veces el experimento para cada valor del número de objetos y de muestras por objeto. En cada una de estas ejecuciones se ha obtenido un valor de Cllr y Min_Cllr, presentando como resultado la media de los valores calculados. Aunque esto permite compensar valores anómalos derivados de la citada generación aleatoria, supone un evidente incremento del tiempo de ejecución de las simulaciones.

Este es uno de los motivos por los que se han establecido los valores arriba indicados de 200 objetos y 10 muestras por objeto. Conceptualmente nada impediría ampliarlos, pero el número de combinaciones se incrementaría notablemente, pudiendo dar lugar a que los tiempos de ejecución de cada experimento fueran inasumibles.

Comparaciones de 1 en 1

El primero de esta serie de experimentos se basa en comparar la muestras de 1 en 1. Las figuras 4.6 y 4.7 muestran los resultados obtenidos, sin y con limitación del Cllr, respectivamente. El número mínimo de muestras por objeto a utilizar es 2.

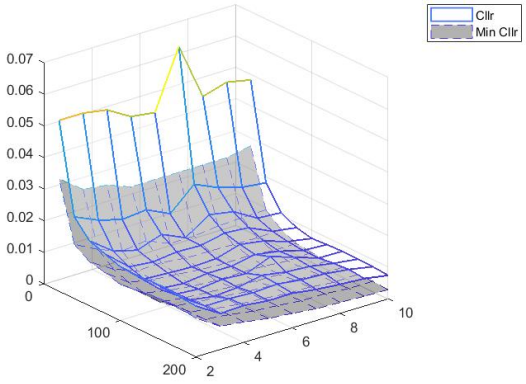

Figura 4.6: Modelo Aitken-Lucy: Comparaciones de 1 en 1 (sin limitación)

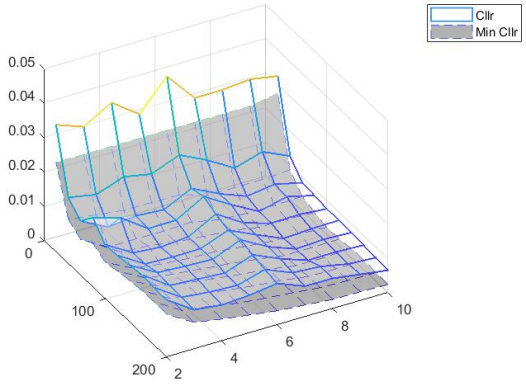

Figura 4.7: Modelo Aitken-Lucy: Comparaciones de 1 en 1 (con limitación)

En este experimento, para las comparaciones dentro del mismo objeto, el valor obtenido en el caso de tomar n muestras será el resultado de comparar la muestra n-sima del objeto con la anterior. Podría pensarse en comparar todos los pares de muestras del objeto pero, por una parte, esto daría lugar a un crecimiento enorme del tiempo de ejecución, que podría hacer inviables los experimentos, y por otra, la limitación realizada permite equilibrar la cantidad de información utilizada en cada experimento.

En las comparaciones entre distintos objetos se compara cada objeto entero (considerando todas sus muestras) con el resto de los objetos.

Como era previsible, la bondad de los resultados (reflejada en la proximidad de los planos) aumenta según crecen el número de objetos y el de muestras por objeto, siendo mucho más sensibles al primero. La aproximación entre los planos no es uniforme, obteniéndose singularidades perfectamente justificables tanto por el origen aleatorio de los datos como por el carácter estadístico del modelo.

Las gráficas también permiten comprobar que el modelo con limitaciones, reflejado en la

segunda de ellas, permite mejorar notablemente los resultados en el caso de contar con pocos objetos, siendo su aportación mucho menor en el caso de un número elevado de objetos, en el que los planos están ya suficientemente próximos en el modelo sin limitación.

Comparaciones de 2 en 2

Los resultados obtenidos al comparar las muestras de 2 en 2, sin y con limitación, respectivamente, se muestran en las figuras 4.8 y 4.9.

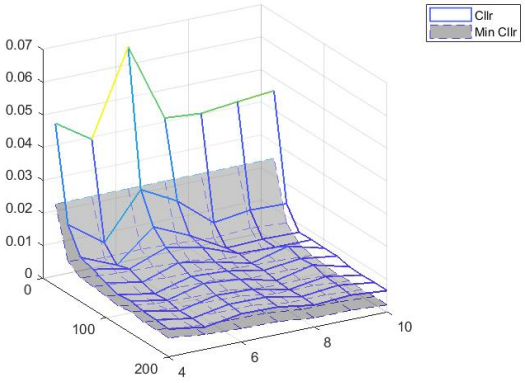

Figura 4.8: Modelo Aitken-Lucy: Comparaciones de 2 en 2 (sin limitación)

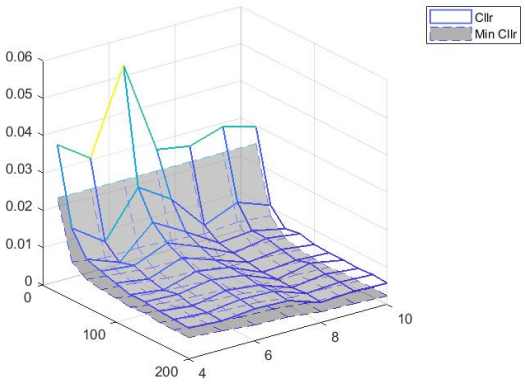

Figura 4.9: Modelo Aitken-Lucy: Comparaciones de 2 en 2 (con limitación)

En las comparaciones dentro del mismo objeto, el valor obtenido en el caso de tomar n muestras será el resultado de comparar la muestra n-sima del objeto y la anterior con las dos previas. El hecho de trabajar con pares de muestras hace que el número mínimo de muestras posibles por objeto sea 4.

En las comparaciones entre distintos objetos se compara cada objeto entero (considerando todos sus pares de muestras) con los del resto de los objetos.

Por una parte, las observaciones realizadas en la comparación de 1 en 1 son plenamente válidas en este caso (mayor influencia del aumento del número de objetos que del número de muestras, pequeñas variaciones aleatorias en datos intermedios, mejores resultados en la prueba con limitación). Cabe una observación adicional: los valores obtenidos tanto para Cllr como para Min_Cllr son menores en la comparación de dos en dos que en la previa (los datos concretos, que

permiten comprobar mejor esta observación, se encuentran en el Anexo 2). No sólo disminuyen ambos valores sino, en general, la diferencia entre ambos.

Comparaciones de 3 en 3

Los resultados obtenidos al comparar las muestras de 3 en 3, sin y con limitación son mostrados en las figuras 4.10 y 4.11, respectivamente.

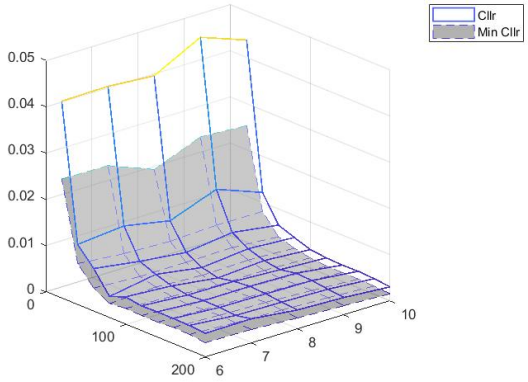

Figura 4.10: Modelo Aitken-Lucy: Comparaciones de 3 en 3 (sin limitación)

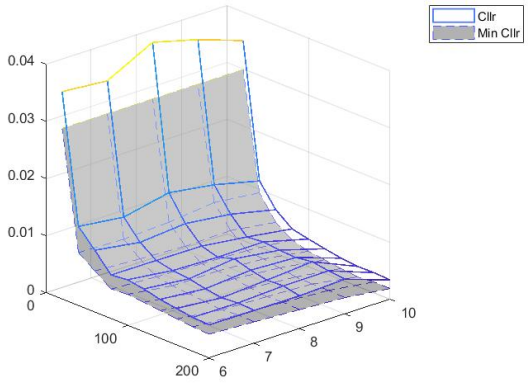

Figura 4.11: Modelo Aitken-Lucy: Comparaciones de 3 en 3 (con limitación)

En las comparaciones dentro del mismo objeto, el valor obtenido en el caso de tomar n muestras será el resultado de comparar la muestra n-sima del objeto y las dos anteriores con las tres que las preceden. El hecho de trabajar con tríos de muestras hace que el número mínimo de muestras posibles por objeto sea 6.

En las comparaciones entre distintos objetos se compara cada objeto entero (considerando todos sus tríos de muestras) con los del resto de los objetos.

Se confirman en este caso las tendencias señaladas en los experimentos anteriores, incluyendo la disminución de los valores de Cllr y Min_Cllr y su diferencia según aumenta el número de muestras.

Comparaciones de 4 en 4

Finalmente, las figuras 4.12 y 4.13 muestran los resultados obtenidos al comparar las muestras de 4 en 4, sin y con limitación, respectivamente.

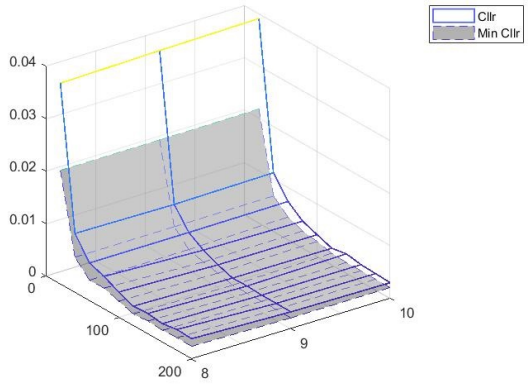

Figura 4.12: Modelo Aitken-Lucy: Comparaciones de 4 en 4 (sin limitación)

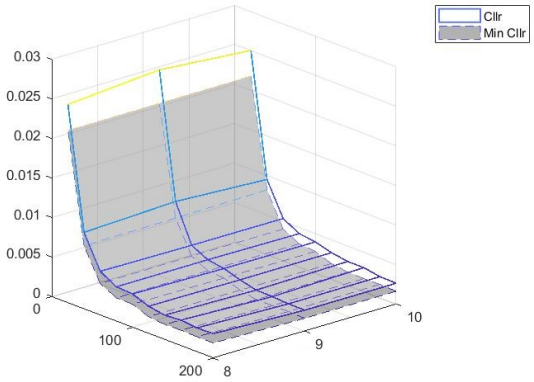

Figura 4.13: Modelo Aitken-Lucy: Comparaciones de 4 en 4 (con limitación)

En las comparaciones dentro del mismo objeto, el valor obtenido en el caso de tomar n muestras será el resultado de comparar la muestra n-sima del objeto y las tres anteriores con las cuatro anteriores. El hecho de trabajar con grupos de cuatro muestras hace que el número mínimo de muestras posibles por objeto sea 8.

En las comparaciones entre distintos objetos se compara cada objeto entero (considerando todos los posibles grupos de 4 muestras) con los del resto de los objetos.

En este caso, siguen siendo válidas las observaciones realizadas en los experimentos anteriores.

4.4. Experimentos realizados con el modelo NFB

Antes de comentar los resultados de los experimentos realizados con el modelo bayesiano completo, conviene hacer algunas consideraciones:

- En general, los experimentos realizados van a ser los mismos que con el modelo anterior.
- En aras de que puedan compararse los resultados obtenidos con ambos modelos, se van a utilizar las mismas bases de datos, es decir:

  - La base de datos que se denominó original, para la medida directa del LR en comparaciones con la misma y distinta fuente, en la que es previsible un resultado menos adecuado el modelo NFB.
  - Para la estimación del rendimiento mediante Cllr y Min_Cllr y el estudio de la influencia del número de objetos y muestras por objeto se utilizará una base de datos ampliada.

- Tal como se ha definido esta versión del modelo bayesiano completo, no es posible la comparación de bloques de objetos, debiéndose realizar la comparación muestra a muestra.

4.4.1. Medida del LR en comparaciones con la misma y distinta fuente

El modelo bayesiano completo descrito en el capítulo 3 ofrecía dos aproximaciones diferentes, en función de que se trabajara con un número limitado o elevado de muestras.

La implementación que se ha codificado utiliza una u otra aproximación en función de las muestras con que se trabaje. En la primera prueba, realizada con la base de datos original, con un número reducido de muestras, la ejecución se basa en las ecuaciones del modelo NFB recogidas en las fórmulas (3.1) y (3.2), es decir con distribuciones *T de student*.

Al igual que en el modelo anterior, el primer experimento que se realizó fue comparar muestras de un mismo objeto y muestras de diferentes objetos para comprobar si los valores obtenidos para el LR correspondían a los esperados.

Para las **comparaciones de la misma fuente**, dado que, según se ha indicado, sólo pueden comparase muestras individuales, y no bloques de muestras, se comparó cada muestra de cada objeto con todas las restantes del mismo objeto.

Como se ha señalado anteriormente, la teoría establece que debería obtenerse un valor de LR > 1. Los resultados obtenidos en la práctica se resumen en el cuadro 4.5.

Valor del LR	**% comparaciones**
< 1	0,48 %
~ 1	0,16 %
1,9 - 10	1,13 %
10 - 100	13,55 %
100 - 1.000	47,10 %
> 1.000	37,58 %

Cuadro 4.5: Modelo NFB: LR para comparación de la misma fuente

Los resultados son bastante similares a los obtenidos en el experimento realizado con el primer modelo: **más del 99 % de las comparaciones generan un LR > 1**. Además, más del 98 % de los casos muestran evidencias claras para determinar que las muestras comparadas proceden de un mismo objeto.

Para el caso de **fuentes diferentes** se realizaron comparaciones individuales ente muestras de diferentes objetos.

Para este tipo de comparaciones se esperaría obtener un valor de LR < 1. En la práctica, los valores obtenidos para el LR fueron los reflejados en el cuadro 4.6.

Valor del LR	**% comparaciones**
»1	9,01 %
~ 1	1,03 %
0,1 - 0,85	3,76 %
0,01 - 0,1	5,63 %
$< 0{,}01$	80,56 %

Cuadro 4.6: Modelo NFB: LR para comparación de distinta fuente

En este caso, **algo más de un 9 % ofrecen un resultado diferente al esperado, es decir, generan un LR > 1**, cuando al comparar muestras procedentes de objetos diferentes el LR obtenido debería ser < 1. Por otro lado, cerca del 90 % de las comparaciones permitirían establecer con claridad que las muestras comparadas proceden de objetos diferentes.

En conclusión, los resultados se ajustan a lo previsto y son similares a los del primer modelo. Sin embargo, **los LR obtenidos en las comparaciones entre muestras de distintos objetos son ligeramente peores en esta primera prueba con el modelo bayesiano completo**.

Al repetir el experimento anterior con la base de datos ampliada, en su ejecución se utiliza la segunda aproximación, en la que se sustituyen las funciones *T de student* por gaussianas.

Los resultados obtenidos están recogidos en los cuadros 4.7 y 4.8.

Valor del LR	**% comparaciones**
< 1	0,20 %
~ 1	0,15 %
1,9 - 10	2,50 %
10 - 100	13,10 %
100 - 1.000	47,90 %
> 1.000	36,15 %

Cuadro 4.7: Modelo NFB (2): LR para comparación de la misma fuente

Valor del LR	**% comparaciones**
»1	3,52 %
~ 1	0,10 %
0,1 - 0,85	0,67 %
0,01 - 0,1	0,48 %
$< 0{,}01$	95,25 %

Cuadro 4.8: Modelo NFB (2): LR para comparación de distinta fuente

En estos cuadros puede observarse que los resultados obtenidos son mejores que los de cualquiera de los modelos y aproximaciones anteriores, lo que podía preverse, ya que se utilizan hipótesis gaussianas con unos datos distribuidos normalmente.

4.4.2. Estimación del rendimiento mediante Cllr y Min_Cllr

Al igual que se hizo en el modelo anterior, se va a medir la bondad de los resultados mediante las curvas ECE. Las figuras 4.14 y 4.15 recogen las curvas obtenidas para el modelo bayesiano completo con la base de datos original y ampliada, respectivamente.

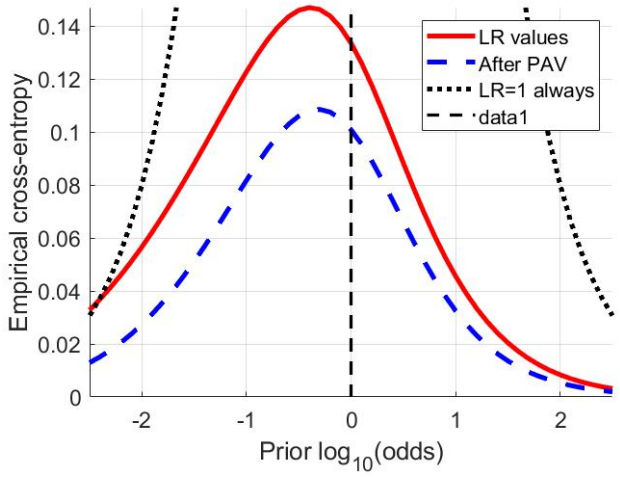

Figura 4.14: Modelo NFB: Curvas ECE de la base de datos original

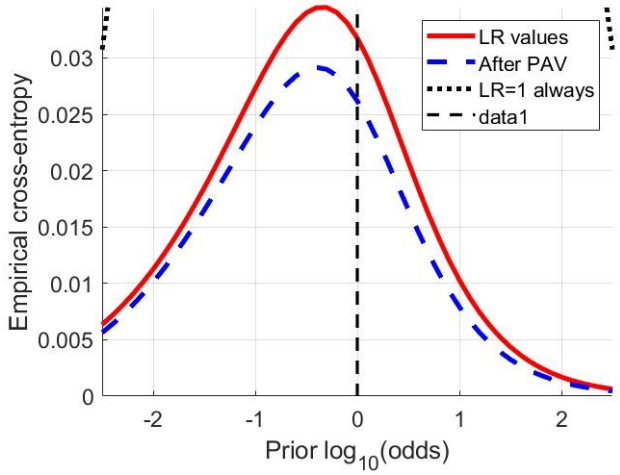

Figura 4.15: Modelo NFB: Curvas ECE de la base de datos ampliada

Como puede observarse en ellas, las curvas empíricas están mucho más próximas a la óptima, lo que refleja una mejora con relación al primer modelo, especialmente en la curva correspondiente a la base de datos ampliada.

Para facilitar la comparación con los resultados obtenidos con el primer modelo, en las figuras 4.16 y 4.17 se recogen las curvas ECE obtenidas para el modelo Aitken-Lucy y el modelo NFB, para las bases de datos original y ampliada, respectivamente.

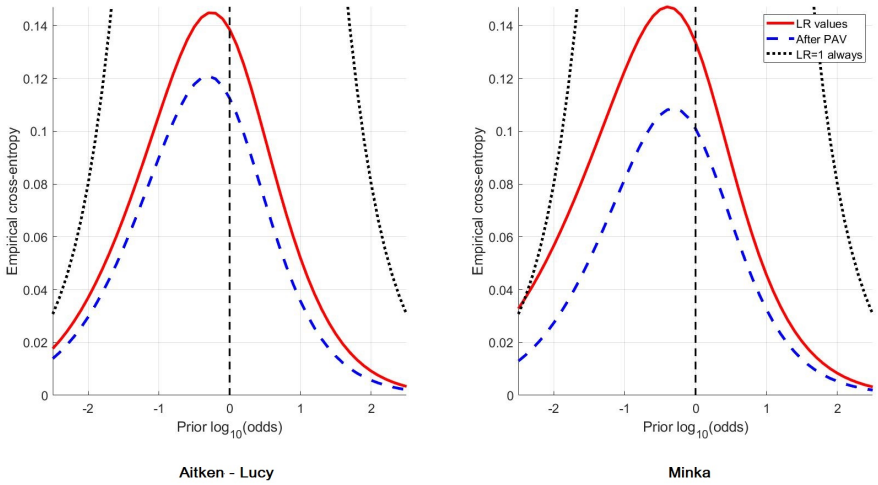

Figura 4.16: Comparación de curvas ECE (base de datos original)

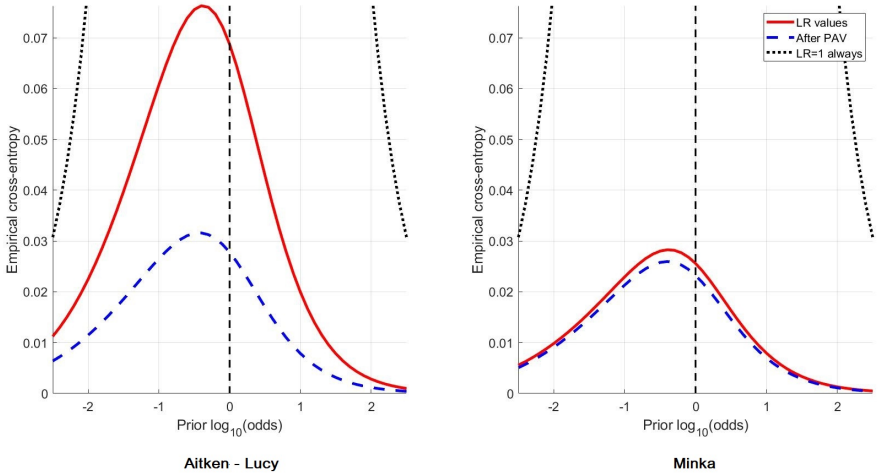

Figura 4.17: Comparación de curvas ECE (base de datos ampliada)

En la primera de ellas, sorprende que el modelo NFB sea comparable al modelo Aitken-Lucy, ya que el primero supone normalidad, siendo los datos claramente no gaussianos. Una posible explicación es que la introducción de la incertidumbre en el modelo NFB compensa esta falta de ajuste de los datos.

En la segunda figura, aunque sea una base de datos simulada, el modelo Minka es extraordinariamente bueno, como puede apreciarse en la proximidad de ambas curvas. El modelo Aitken-Lucy debería tener un buen comportamiento (ya que una distribución kernel se ajusta bien a una gaussiana), pero su falta de ajuste en este caso puede deberse a la falta de datos, que el modelo no puede asumir.

4.4.2.1. Resultados en función del número de objetos y muestras

A continuación, se realizó un experimento similar al del modelo anterior para analizar la influencia que tiene sobre los resultados la variación en el número de objetos y muestras.

Igual que en las pruebas realizadas con el modelo de Aitken-Lucy, se requiere un número elevado de objetos y muestras para tener un margen de variación, por lo que se trabajó con la base de datos ampliada. Sin embargo, dado que tal como se ha implementado el modelo NFB sólo admite comparaciones 1 a 1, solo se realizarán este tipo de comparaciones.

La figura 4.18 recoge los resultados obtenidos en este experimento. Se ha aplicado directamente el modelo, sin acotar los resultados obtenidos para el LR entre sus valores mínimo y máximo teóricos (se trata de lo que se denominó para el primer modelo comparación sin limitación de LR).

Para reducir la aleatoriedad de los resultados obtenidos en función de la base de datos ampliada generada, se ha repetido varias veces el experimento y lo que se presenta son los valores medios.

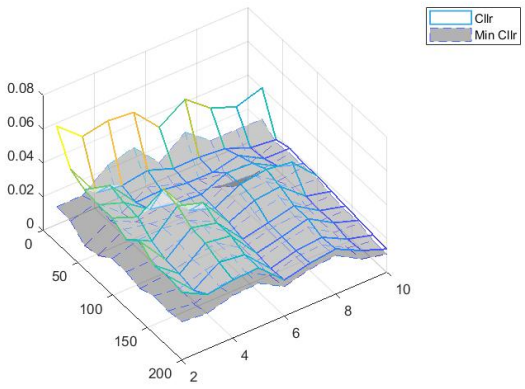

Figura 4.18: Modelo NFB: Comparaciones sin limitación

A la hora de analizar esa figura, no se puede perder de vista que los resultados obtenidos con el modelo NFB y una base de datos ampliada son en general bastante mejores que los obtenidos con el modelo anterior. Es decir, las diferencias entre las curvas empírica y óptima son bastante limitadas.

Por este motivo, la distancia entre los planos de Cllr y Min_Cllr es menor, por lo que resalta cualquier pequeña variación debida a la aleatoriedad inherente a las pruebas. No es de extrañar, por tanto, que aparezcan en las figuras pequeñas irregularidades.

Puede observarse en la imagen que los resultados obtenidos mejoran inicialmente al aumentar el número de objetos y muestras, pero convergen enseguida a unos valores más o menos contantes. Lo que esto refleja, en definitiva, es la mayor estabilidad del modelo NFB frente al modelo de Aitken-Lucy.

4.4.2.2. Comparación entre modelos

Para finalizar el estudio, se repitió exactamente este experimento utilizando el modelo Aitken-Lucy, con los mismos datos utilizados en esta comparación para el modelo NFB. Dado que la implementación desarrollada para el modelo NFB sólo admite comparaciones 1 a 1, se utilizó este tipo de comparaciones con el modelo anterior, para el que se ejecutó tanto su versión sin limitación como su versión con limitación.

La figura 4.19 muestra los resultados obtenidos en cada uno de estos tres casos. Los resultados numéricos de estos experimentos están recogidos en el anexo 2.

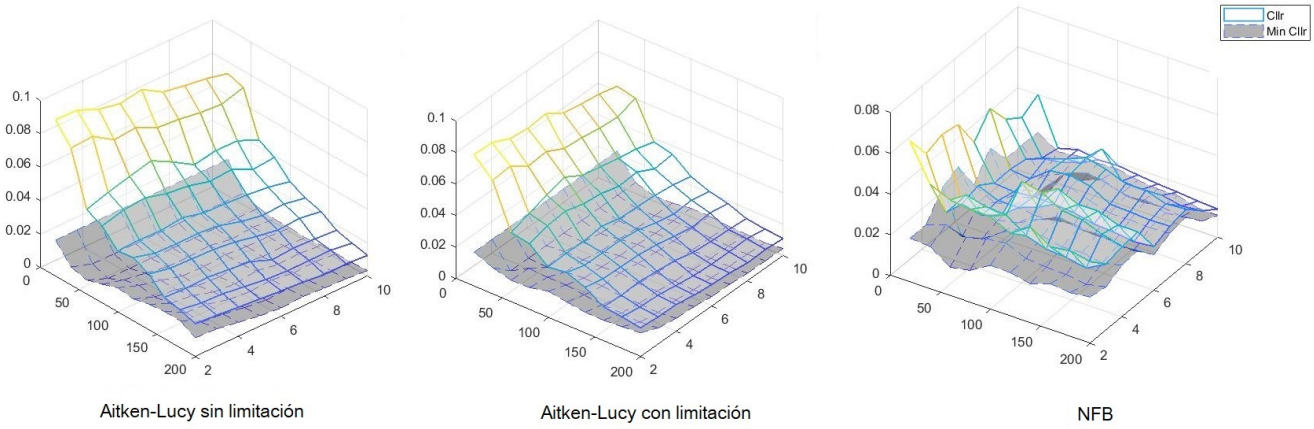

Figura 4.19: Comparación entre modelos

Como puede verse, tal y como se ha ido observando en las diferentes pruebas, los resultados del modelo NFB con la base de datos ampliada mejoran claramente los ofrecidos por el modelo de Aitken-Lucy, en sus dos versiones (con o sin limitación).

5

Conclusiones y trabajo futuro

El propósito de este Trabajo de Fin de Grado, según se planteó en la introducción de la memoria, es comprobar la utilidad de la estadística (en particular, de la estadística bayesiana) en el ámbito de la ciencia forense, proporcionando modelos para resolver el problema de comparación de muestras de cristales.

Pretende ir más allá de implementar un modelo conocido como es el de Aitken-Lucy, ya que propone un modelo nuevo que supere alguna de sus limitaciones de ese primer modelo.

La aproximación elegida para ambos modelos se ha basado en la utilización del *likelihood ratio* (LR) como valor numérico que permita resolver dicho problema y se han seleccionado dos modelos matemáticos para asignar ese valor.

Antes de comentar y valorar los resultados, se subrayarán algunos hechos que los condicionan:

- Se dispone de pocos datos reales con los que calcular y asignar valor para el LR. Por ese motivo, a partir de ellos se han generado aleatoriamente unos datos de prueba para realizar los sucesivos experimentos. A la hora de justificar esa ampliación, ya se advirtió de que puede introducir sesgos que distorsionen ligeramente los resultados.
- En los modelos seleccionados se establecen hipótesis sobre la distribución de los datos. En la medida en que esas hipótesis correspondan con la realidad el modelo será más o menos adecuado.
- Se ha observado empíricamente que los resultados son mejores cuando aumenta tanto el número de objetos que componen la base de datos como el de muestras por objeto. La cantidad de comparaciones a realizar crece con el cuadrado de dicho número; según aumenta el tamaño de la base de datos, el tiempo de ejecución crece conforme a dicho cuadrado, dando lugar a unos tiempos muy elevados. Esto impone una limitación en la práctica, al menos en lo que se refiere a los experimentos realizados en el marco de este proyecto.
- A este respecto, hay que recordar que en los experimentos realizados con la base de datos ampliada se ha repetido cada prueba varias veces, para ofrecer como resultado la media de las sucesivas ejecuciones. De esta manera, se pretende compensar los posibles errores derivados de la aleatoriedad causada por esa ampliación, pero multiplica el tiempo de ejecución por el número de repeticiones.

Como se señaló en el capítulo 4, los resultados con ambos modelos son razonablemente satisfactorios, debiendo subrayarse que:

- Los valores obtenidos para el LR, tanto cuando se hacen comparaciones dentro del mismo objeto como cuando son muestras de diferentes objetos responden a lo esperado en un porcentaje elevado de casos.
- Además, los valores próximos a la frontera (LR = 1) son bastante escasos, es decir, el modelo se decanta claramente sobre si las muestras pertenecen al mismo objeto o no.
- Sin embargo, de cara a su utilización en la práctica como herramienta forense, sería deseable un porcentaje de aciertos todavía mayor y reducir los casos situados en la frontera, pues no ayudan a reducir la incertidumbre.
- Sin embargo, se observa una diferencia entre resultados que se obtienen cuando se realiza el experimento con la base de datos original y cuando se realiza con la base de datos ampliada, generado aleatoriamente a partir de la anterior.
- El análisis realizado mediante comparación de Cllr y Min_Cllr muestra unos mejores resultados cuando aumenta el número de objetos de que se dispone; es menos sensible al aumento del número de muestras por objeto.
- La introducción de límites al valor del LR (dentro del margen entre el mínimo y el máximo teórico) también supone una ligera mejoría en los resultados obtenidos.
- Los resultados obtenidos con el modelo bayesiano completo (NFB) son mejores que los del primer modelo (Aitken-Lucy) en general. Este hecho es particularmente acusado cuando se dispone de poca información (pocos objetos y muestras por objeto); en consecuencia, la variabilidad de los resultados al variar el número de objetos y muestras es menor o, dicho con otras palabras, tiene una mayor estabilidad.
- La diferencia de resultados entre modelos también es más acusada al utilizar la base de datos ampliada, cuya distribución es gaussiana, dado que ese tipo de distribución es una de las hipótesis de las que parte el modelo NFB.

Finalmente, se proponen diferentes líneas de trabajo futuro para ampliar el estudio de este proyecto:

- Se considera muy adecuado repetir los experimentos con bases de datos de mayor tamaño que contengan datos reales, que permitan confirmar las tendencias apuntadas con los datos utilizados en este proyecto.
- En su defecto, para refrendar los resultados obtenidos y eliminar sesgos derivados de la aleatoriedad convendría repetir los experimentos un número elevado de veces, lo que supone un problema de potencia de cálculo y/o tiempo de ejecución.
- También convendría utilizar bases de datos reales que contengan un mayor número de características a comparar. Existe una certeza de que al aumentar el número de dimensiones los resultados obtenidos se degradan considerablemente.
- Por último, se podrían modificar las hipótesis en las que se basan ambos modelos y comprobar su efecto sobre los resultados obtenidos. Por ejemplo, igual que en el modelo de Aitken-Lucy se han sustituido distribuciones gaussianas por *kernel* podría realizarse algo análogo con el modelo NFB.

En definitiva, se han alcanzado satisfactoriamente todos los objetivos que se habían propuesto en el capítulo 1:

1. Se ha realizado la formulación matemática de dos modelos aplicables a la ciencia forense; uno de ellos de uso habitual y otro modelo original.
2. Se han implementado los modelos propuestos.
3. Se han analizado los resultados de las simulaciones realizadas, comparando su validez y robustez ante la falta de datos, reflejándose los mejores resultados del modelo nuevo.
4. Se han identificado limitaciones de los modelos.
5. Se han realizado propuestas de mejora.

Glosario y acrónimos

- **Bayesiano:** Relativo al teorema de Bayes y su aplicación.
- **Between-source variation (Comparación de distinta fuente)**: Variación existente entre observaciones procedentes de objetos diferentes.
- **Ciencia forense:** Conjunto de disciplinas científicas que ayudan a la policía y la justicia a determinar las circunstancias exactas de la comisión de una infracción y a identificar a sus autores, permitiendo la aplicación de prácticas científicas dentro de un proceso legal.
- **Cllr (*Log-Likelihood Ratio Cost*):** Puede considerarse como una medida estándar de la precisión de un sistema que genera LRs. En las curvas ECE, es el valor de la curva empírica (obtenida a partir de unos LRs) cuando corta el eje vertical.
- **Curvas ECE (*Empirical Cross-Entropy*)**: Tipo de curvas que permiten una representación compacta de la entropía cruzada.
- **Diagrama de dispersión *(Scatter plot*):** Es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos. Uno de los aspectos más poderosos de un gráfico de dispersión es su capacidad para mostrar las relaciones no lineales entre las variables. Además, si los datos son representados por un modelo de mezcla de relaciones simples, estas relaciones son visualmente evidentes como patrones superpuestos.
- **Entropía empírica cruzada:** Métrica que puede utilizarse para reflejar la precisión de los pronósticos probabilísticos. Está estrechamente vinculada con la estimación por máxima verosimilitud.
- **Estadística bayesiana:** Es un subconjunto del campo de la estadística en la que la evidencia sobre el verdadero estado del mundo se expresa en términos de grados de creencia o, más específicamente, las probabilidades bayesianas. Ésta es sólo una de las muchas interpretaciones existentes de la probabilidad, ya que hay otras técnicas estadísticas que no se basan en "grados de creencia".

  Una de sus ideas fundamentales es que la probabilidad es una opinión organizada, y que la inferencia a partir de los datos no es más que la revisión de esa opinión a la luz de nueva información relevante.
- **Falso negativo:** En un estudio de investigación experimental, se refiere al **error de tipo II**, es decir, aquel que se comete cuando el investigador no rechaza la hipótesis nula siendo ésta falsa en la población. Es equivalente a la probabilidad de un resultado falso negativo, ya que el investigador llega a la conclusión de que ha sido incapaz de encontrar una diferencia que existe en la realidad.
- **Falso positivo:** En un estudio de investigación experimental, se refiere al **error de tipo I**, es decir, al que se comete cuando el investigador rechaza la hipótesis nula siendo ésta verdadera en la población. En este caso, el investigador llega a la conclusión de que existe una diferencia entre las hipótesis cuando en realidad no existe.

- **LR (*likelihood ratio*):** El término, comúnmente utilizado en inglés, podría traducirse como cociente, índice, razón o ratio de probabilidades o verosimilitudes. Se define como la razón entre la posibilidad de tener muestras procedentes de un mismo objeto frente a la posibilidad de que procedan de objetos diferentes.
- **Min_Cllr:** Valor mínimo del Cllr. En las curvas ECE, es el valor de la curva óptima cuando corta el eje vertical.
- **PDF (*Probability Density Function*)**: Término que hace referencia a la función de densidad de una determinada distribución..
- **Probabilidad a posteriori (*posterior probability*):** En la estadística Bayesiana, la probabilidad a posteriori de un evento aleatorio es la probabilidad condicional que es asignada después de que la evidencia es tomada en cuenta.
- **Probabilidad a priori (*prior probability*):** En la estadística Bayesiana, una distribución de probabilidad a priori de una cantidad p desconocida es la distribución de probabilidad que expresa alguna incertidumbre acerca de p antes de tomar en cuenta los "datos".
- **Within-source variation (Comparación de la misma fuente)**: Variación existente entre observaciones procedentes de un mismo objeto.

Bibliografía

[1] Daniel Ramos, Javier Franco Pedroso, Alicia Lozano Díez y Joaquín González Rodríguez; Escuela Politécnica Superior (Universidad Autónoma de Madrid). Deconstructing Cross-Entropy for Probabilistic Binary Classifiers. Enviado a Entropy, Febrero 2018.

[2] C.G.G.Aitken y D.Lucy, Universidad de Edimburgo (Reino Unido). Evaluation of Trace Evidence in the Form of Multivariate Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 53(1):109–122, 2004.

[3] J.M. Bernardo, Facultad de Matemáticas (Universidad de Valencia). *Probability and Statistics (capítulo sobre 'Bayesian Statistics')*. UNESCO, Oxford (United Kingdom), 2003. Versión actualizada y revisada.

[4] Niko Brümmer, AGNITIO LABS, África del Sur. Fully Bayesian Score Calibration assuming Gaussian Distributions, Mayo 2011.

[5] M. Cristina Casado Fernández. *Manual básico de MATLAB*. Servicios Informáticos U.C.M, Apoyo a Investigación y Docencia, Madrid. Publicado en: http://webs.ucm.es/centros/cont/descargas/documento11541.pdf.

[6] Richard O. Duda, Peter E. Hart y David G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2ª edition, 2001. ISBN 9780471056690.

[7] Javier Franco Pedroso, Daniel Ramos, Joaquín González Rodríguez, Universidad Autónoma de Madrid. Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. *PLOS One*, Febrero 2016. DOI:10.1371/journal.pone.0149958.

[8] Rudolf Haraksim, Daniel Ramos, Didier Meuwly. Validation of likelihood ratio methods for forensic evidence evaluation handling multimodal score distributions. *IET Biometrics*, Noviembre 2016. 61-69, ISSN 2047-4938.

[9] David J.C. MacKay, Universidad de Cambridge (Reino Unido). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Marzo 2005. Versión 7.2. ISBN 9780521642989.

[10] Thomas P. Minka. Inferring a Gaussian distribution. Technical report, 1998 (revisado en 2001).

[11] Geoffrey Stewart Morrison, Tharmarajah Thiruvaran y Julien Epps. Estimating the Precision of the Likelihood-Ratio Output of a Forensic-Voice-Comparison System. Technical report, Odissey, Brno, República Checa, Julio 2010. The Speaker and Language Recognition Workshop.

[12] Cristhian Müller (Ed.). *Speaker Classification I. Fundamentals, Features and Methods*. Springer, Berkeley, USA, 2007. Lecture Notes in Artificial Intelligence 4343. ISBN9783540741862.

[13] Regina Nuzzo. Statistical Errors. *Nature*, 506:150–152, Febrero 2014. Macmillan Publishers Limited.

[14] Tony O'Hagan. Dicing with the unknown. *Significance*, Septiembre 2004. 132-133.

[15] Javier Ramírez Pérez de Inestrosa. *Tutorial de R.* Universidad de Granada, Departamento de Teoría de la Señal, Telemática y Comunicaciones, Granada. Publicado en: http://www.ugr.es/∼javierrp/master_files/Tutorial%20de%20R.pdf.

[16] Daniel Ramos Castro. ECE plot. Software to draw ECE plots and derived representations. Disponible en http://arantxa.ii.uam.es/∼dramos/software.html, 9 2009. Freely Available Software in $Matlab^{TM}$.

[17] Daniel Ramos Castro. *Evaluación de la evidencia forense utilizando sistemas automáticos de reconocimiento de locutor.* PhD thesis, Universidad Autónoma de Madrid, Madrid, Noviembre 2007.

[18] Ronald L. Wasserstein y Nicole A. Lazar. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, 2016. DOI: 10.1080/00031305.2016.1154108.

[19] Bayes' theorem. *Publicado en Wikipedia.* https://simple.wikipedia.org/wiki/Bayes

[20] Likelihood ratios in diagnostic testing. *Publicado en Wikipedia.* https://en.wikipedia.org/wiki/Likelihood_ratios_in_diagnostic_testing.

[21] Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos y Colin Aitken. *Statistical Analysis in Forensic Science. Evidential Value of Multivariate Physicochemical Data.* John Wiley & Sons, Ltd., West Sussex (Reino Unido), 2014. ISBN 9780470972106.

Anexo 1: Distribuciones de probabilidad

- **Distribución gaussiana multivariable**

  Se caracteriza por la siguiente función de densidad:

$$N(x, \mu, \Sigma) = (2\pi)^{-d/2} \left|\Sigma\right| - 1/2 \quad exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

- **Distribución Gamma**

  Se caracteriza por la siguiente función de densidad:

$$f(x, \tau|\mu, \lambda, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}} \tau^{\alpha - \frac{1}{2}} e^{-\beta\tau} e^{-\frac{\lambda\tau(x-\mu)^2}{2}}$$

- **Distribución de Bernoulli**

  Es una distribución de probabilidad discreta, que toma valor 1 para la probabilidad de éxito (p) y valor 0 para la probabilidad de fracaso ($q = 1 - p$).

  Si X es una variable aleatoria que mide el "número de éxitos", y se realiza un único experimento con dos posibles resultados (éxito o fracaso), se dice que la variable aleatoria X se distribuye como una Bernoulli de parámetro p, cuya función de probabilidad viene definida por:

$$f(x) = p^x (1-p)^{1-x}$$

- **Distribución T de student**

  Un vector aleatorio x de longitud d se dice que tiene una distribución T de student con parámetros m, V y n si la densidad de x es

$$p(x) \sim \tau(m, V, n) = \frac{\Gamma(n/2)}{\Gamma((n-d)/2)} \left|\pi V\right|^{-1/2} \left((x-m)^T V^{-1}(x-m) + 1\right)^{-n/2}$$

  En el caso particular unidimensional,

$$p(x) \sim \tau(m, v, n) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \frac{1}{\sqrt{\pi v}} \left(\frac{(x-m)^2}{v} + 1\right)^{-n/2}$$

- **Distribución Kernel**

  La distribución kernel se utiliza cuando no existe una distribución paramétrica que se ajuste correctamente a los datos, y se define mediante una función de alisado (*smoothing function*) y un ancho de banda.

  Se caracteriza por la siguiente estimación de la función de densidad:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

  siendo $x_1, \ldots, x_n$ muestras aleatorias de una distribución desconocida, n el tamaño de la muestra, K la función de alisado y h el ancho de banda.

Anexo 2: Resultados de experimentos

En los siguientes cuadros se representan los valores obtenidos para Cllr y Min_Cllr en función del número de objetos y el número de muestras (M).

RESULTADOS OBTENIDOS CON EL MODELO AITKEN-LUCY

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00740	0,00780	0,00787	0,00803	0,00783	0,00747	0,00857	0,01100	0,01813	0,04833
9	0,00868	0,00925	0,00885	0,00820	0,00828	0,00878	0,00970	0,01153	0,01900	0,04995
8	0,00918	0,01018	0,01020	0,01078	0,01035	0,01130	0,01315	0,01593	0,02120	0,04775
7	0,00988	0,01055	0,01070	0,01328	0,01160	0,01250	0,01378	0,01725	0,02460	0,06585
6	0,01098	0,01153	0,01200	0,01025	0,01023	0,01028	0,01215	0,01230	0,01755	0,04725
5	0,01093	0,01113	0,01048	0,01075	0,01028	0,01000	0,01243	0,01565	0,02115	0,04830
4	0,01178	0,01155	0,00993	0,00993	0,00938	0,01033	0,01155	0,01298	0,02038	0,05273
3	0,01150	0,01153	0,01115	0,01058	0,01003	0,01083	0,01235	0,01495	0,02245	0,05400
2	0,01610	0,01650	0,01718	0,01708	0,01668	0,01703	0,01878	0,02050	0,02583	0,05408

Cuadro 1: Modelo Aitken-Lucy: Cllr (Comparación de 1 en 1, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00287	0,00323	0,00307	0,00293	0,00263	0,00303	0,00357	0,00567	0,00957	0,02750
9	0,00308	0,00323	0,00285	0,00308	0,00270	0,00268	0,00358	0,00535	0,00938	0,02608
8	0,00345	0,00390	0,00365	0,00360	0,00300	0,00333	0,00433	0,00553	0,00908	0,02608
7	0,00378	0,00393	0,00383	0,00403	0,00323	0,00353	0,00420	0,00573	0,00818	0,02608
6	0,00460	0,00473	0,00458	0,00450	0,00373	0,00338	0,00453	0,00585	0,00818	0,02608
5	0,00533	0,00530	0,00508	0,00533	0,00428	0,00390	0,00535	0,00675	0,01133	0,02608
4	0,00653	0,00650	0,00608	0,00603	0,00478	0,00490	0,00603	0,00643	0,01040	0,02895
3	0,00748	0,00738	0,00740	0,00690	0,00590	0,00638	0,00743	0,00928	0,01318	0,02993
2	0,01230	0,01260	0,01303	0,01283	0,01195	0,01145	0,01285	0,01393	0,01713	0,03553

Cuadro 2: Modelo Aitken-Lucy: Min_Cllr (Comparación de 1 en 1, sin limitación de LR)

	Número de objetos									
M										
10	0,00653	0,00680	0,00705	0,00733	0,00753	0,00747	0,00797	0,00827	0,01267	0,03253
9	0,00770	0,00730	0,00743	0,00768	0,00738	0,00815	0,00993	0,01018	0,01625	0,03390
8	0,00885	0,00890	0,00735	0,00725	0,00690	0,00798	0,00800	0,00903	0,01385	0,03318
7	0,00893	0,00943	0,01010	0,01020	0,01033	0,01035	0,01228	0,01208	0,01853	0,03315
6	0,01208	0,01270	0,01360	0,01460	0,01398	0,01530	0,01698	0,01598	0,02233	0,04178
5	0,01098	0,01043	0,01065	0,01125	0,01098	0,01103	0,01248	0,01210	0,01913	0,03315
4	0,01048	0,01085	0,01118	0,01158	0,01085	0,01035	0,01238	0,01270	0,02065	0,03883
3	0,01143	0,01165	0,01170	0,01208	0,01118	0,01105	0,01788	0,01183	0,01780	0,03413
2	0,01665	0,01680	0,01713	0,01653	0,01603	0,01575	0,02025	0,01585	0,01935	0,03705

Cuadro 3: Modelo Aitken-Lucy: Cllr (Comparación de 1 en 1, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00243	0,00250	0,00267	0,00277	0,00207	0,00210	0,00307	0,00430	0,00863	0,02750
9	0,00263	0,00250	0,00250	0,00235	0,00190	0,00250	0,00338	0,00410	0,00818	0,02608
8	0,00273	0,00280	0,00263	0,00245	0,00195	0,00273	0,00340	0,00408	0,00815	0,02608
7	0,00348	0,00363	0,00380	0,00365	0,00315	0,00350	0,00480	0,00453	0,00818	0,02608
6	0,00413	0,00435	0,00458	0,00455	0,00420	0,00445	0,00545	0,00625	0,01073	0,02608
5	0,00525	0,00528	0,00558	0,00560	0,00505	0,00475	0,00615	0,00593	0,01085	0,02608
4	0,00640	0,00675	0,00688	0,00665	0,00590	0,00543	0,00735	0,00600	0,01093	0,02608
3	0,00773	0,00788	0,00800	0,00813	0,00713	0,00663	0,00790	0,00600	0,01195	0,02608
2	0,01250	0,01265	0,01263	0,01208	0,01148	0,01068	0,01285	0,01055	0,01295	0,02608

Cuadro 4: Modelo Aitken-Lucy: Min_Cllr (Comparación de 1 en 1, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00648	0,00654	0,00638	0,00654	0,00740	0,00742	0,00756	0,00742	0,01378	0,04592
9	0,00742	0,00800	0,00826	0,00940	0,00890	0,01028	0,00760	0,00766	0,01414	0,04486
8	0,00668	0,00686	0,00658	0,00696	0,00766	0,00738	0,00796	0,00722	0,01258	0,04362
7	0,00842	0,00898	0,00898	0,00966	0,00976	0,01010	0,01216	0,01336	0,02154	0,04450
6	0,00864	0,00912	0,00946	0,01066	0,01038	0,01218	0,01436	0,01834	0,02734	0,06856
5	0,00688	0,00712	0,00700	0,00784	0,00746	0,00850	0,01022	0,00894	0,01344	0,04262
4	0,00858	0,00900	0,00922	0,00994	0,00944	0,01072	0,01238	0,01362	0,02128	0,04984

Cuadro 5: Modelo Aitken-Lucy: Cllr (Comparación de 2 en 2, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00210	0,00228	0,00200	0,00226	0,00194	0,00234	0,00278	0,00396	0,00790	0,02522
9	0,00278	0,00292	0,00266	0,00308	0,00246	0,00300	0,00336	0,00430	0,00862	0,02522
8	0,00276	0,00296	0,00262	0,00300	0,00236	0,00290	0,00362	0,00396	0,00790	0,02522
7	0,00396	0,00410	0,00408	0,00454	0,00386	0,00458	0,00528	0,00566	0,00860	0,02522
6	0,00412	0,00424	0,00412	0,00482	0,00396	0,00462	0,00516	0,00612	0,00908	0,02522
5	0,00462	0,00470	0,00438	0,00510	0,00436	0,00486	0,00608	0,00644	0,00878	0,02522
4	0,00590	0,00602	0,00596	0,00668	0,00602	0,00662	0,00714	0,00734	0,01008	0,02512

Cuadro 6: Modelo Aitken-Lucy: Min_Cllr (Comparación de 2 en 2, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00574	0,00604	0,00570	0,00594	0,00590	0,00616	0,00706	0,00640	0,01070	0,02966
9	0,00604	0,00646	0,00660	0,00756	0,00634	0,00724	0,00824	0,00826	0,01388	0,03218
8	0,00548	0,00586	0,00544	0,00568	0,00554	0,00624	0,00674	0,00744	0,01020	0,02938
7	0,00834	0,00886	0,00852	0,00920	0,00930	0,00930	0,01084	0,01288	0,01966	0,03092
6	0,00862	0,00910	0,00948	0,01040	0,01066	0,01236	0,01430	0,01784	0,02524	0,05600
5	0,00778	0,00804	0,00782	0,00894	0,00816	0,00908	0,01070	0,00954	0,01308	0,03346
4	0,00834	0,00864	0,00884	0,00934	0,00930	0,01040	0,01162	0,01284	0,01916	0,03952

Cuadro 7: Modelo Aitken-Lucy: Cllr (Comparación de 2 en 2, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00234	0,00256	0,00234	0,00248	0,00206	0,00234	0,00278	0,00396	0,00790	0,02522
9	0,00270	0,00290	0,00270	0,00300	0,00246	0,00286	0,00336	0,00430	0,00862	0,02522
8	0,00296	0,00326	0,00284	0,00302	0,00236	0,00276	0,00340	0,00396	0,00790	0,02522
7	0,00400	0,00416	0,00398	0,00436	0,00368	0,00414	0,00462	0,00566	0,00860	0,02522
6	0,00426	0,00442	0,00434	0,00484	0,00408	0,00458	0,00496	0,00580	0,00908	0,02522
5	0,00514	0,00516	0,00490	0,00550	0,00458	0,00502	0,00600	0,00654	0,00878	0,02522
4	0,00578	0,00584	0,00584	0,00640	0,00596	0,00644	0,00672	0,00742	0,01012	0,02522

Cuadro 8: Modelo Aitken-Lucy: Min_Cllr (Comparación de 2 en 2, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00303	0,00338	0,00358	0,00350	0,00365	0,00403	0,00523	0,00650	0,01215	0,04368
9	0,00378	0,00415	0,00460	0,00478	0,00413	0,00470	0,00585	0,00795	0,01588	0,04728
8	0,00413	0,00438	0,00478	0,00500	0,00465	0,00400	0,00483	0,00668	0,01213	0,04200
7	0,00525	0,00455	0,00488	0,00493	0,00475	0,00530	0,00638	0,00780	0,01403	0,04278
6	0,00515	0,00543	0,00553	0,00555	0,00543	0,00598	0,00465	0,00928	0,01308	0,04260

Cuadro 9: Modelo Aitken-Lucy: Cllr (Comparación de 3 en 3, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00155	0,00178	0,00190	0,00193	0,00188	0,00225	0,00293	0,00400	0,00805	0,02520
9	0,00200	0,00233	0,00248	0,00240	0,00225	0,00258	0,00343	0,00470	0,00813	0,02578
8	0,00240	0,00270	0,00295	0,00278	0,00243	0,00240	0,00315	0,00455	0,00768	0,02180
7	0,00295	0,00300	0,00320	0,00298	0,00278	0,00305	0,00365	0,00453	0,00805	0,02565
6	0,00313	0,00350	0,00363	0,00335	0,00338	0,00350	0,00315	0,00505	0,00895	0,02580

Cuadro 10: Modelo Aitken-Lucy: Min_Cllr (Comparación de 3 en 3, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00338	0,00350	0,00345	0,00375	0,00425	0,00480	0,00543	0,00765	0,01155	0,03495
9	0,00608	0,00640	0,00683	0,00643	0,00658	0,00688	0,00660	0,00945	0,01345	0,03770
8	0,00575	0,00633	0,00620	0,00680	0,00765	0,00678	0,00758	0,01038	0,01465	0,03980
7	0,00578	0,00593	0,00568	0,00613	0,00673	0,00703	0,00765	0,00968	0,01295	0,03565
6	0,00583	0,00598	0,00640	0,00685	0,00753	0,00800	0,00768	0,01088	0,01390	0,03633

Cuadro 11: Modelo Aitken-Lucy: Cllr (Comparación de 3 en 3, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00188	0,00188	0,00163	0,00193	0,00233	0,00268	0,00353	0,00545	0,00948	0,02990
9	0,00285	0,00298	0,00295	0,00305	0,00348	0,00390	0,00423	0,00660	0,01035	0,02990
8	0,00333	0,00348	0,00333	0,00368	0,00418	0,00438	0,00450	0,00643	0,00948	0,02990
7	0,00378	0,00385	0,00358	0,00388	0,00410	0,00438	0,00480	0,00680	0,00908	0,02990
6	0,00418	0,00433	0,00460	0,00490	0,00520	0,00545	0,00523	0,00725	0,00933	0,02990

Cuadro 12: Modelo Aitken-Lucy: Min_Cllr (Comparación de 3 en 3, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00293	0,00308	0,00320	0,00368	0,00315	0,00415	0,00515	0,00733	0,01123	0,03905
9	0,00318	0,00320	0,00320	0,00335	0,00328	0,00423	0,00545	0,00723	0,01100	0,03875
8	0,00388	0,00405	0,00358	0,00393	0,00373	0,00485	0,00625	0,00723	0,01128	0,03835

Cuadro 13: Modelo Aitken-Lucy: Cllr (Comparación de 4 en 4, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00201	0,00201	0,00205	0,00203	0,00220	0,00288	0,00328	0,00438	0,00680	0,02180
9	0,00203	0,00205	0,00203	0,00200	0,00193	0,00265	0,00318	0,00398	0,00675	0,02173
8	0,00243	0,00248	0,00230	0,00250	0,00228	0,00313	0,00403	0,00395	0,00675	0,02173

Cuadro 14: Modelo Aitken-Lucy: Min_Cllr (Comparación de 4 en 4, sin limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00260	0,00265	0,00301	0,00301	0,00335	0,00393	0,00420	0,00533	0,00950	0,02505
9	0,00263	0,00265	0,00278	0,00283	0,00335	0,00400	0,00433	0,00538	0,01015	0,02603
8	0,00310	0,00305	0,00325	0,00310	0,00345	0,00423	0,00435	0,00553	0,00968	0,02510

Cuadro 15: Modelo Aitken-Lucy: Cllr (Comparación de 4 en 4, con limitación de LR)

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00159	0,00168	0,00163	0,00168	0,00198	0,00273	0,00292	0,00401	0,00818	0,02180
9	0,00160	0,00168	0,00165	0,00163	0,00205	0,00280	0,00293	0,00403	0,00805	0,02180
8	0,00193	0,00193	0,00198	0,00193	0,00213	0,00293	0,00295	0,00398	0,00798	0,02180

Cuadro 16: Modelo Aitken-Lucy: Min_Cllr (Comparación de 4 en 4, con limitación de LR)

RESULTADOS OBTENIDOS EN LA COMPARACIÓN ENTRE MODELOS

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,01240	0,01430	0,01670	0,02260	0,02530	0,02980	0,03900	0,04410	0,07010	0,07310
9	0,01340	0,01620	0,01760	0,02460	0,02780	0,03250	0,03770	0,04960	0,06620	0,07610
8	0,01300	0,01600	0,01810	0,02360	0,02690	0,03390	0,04370	0,04970	0,06710	0,07790
7	0,01450	0,01590	0,01710	0,02580	0,02930	0,03470	0,04270	0,04730	0,06910	0,08020
6	0,01530	0,01670	0,02210	0,02850	0,03290	0,03520	0,04600	0,05500	0,07020	0,08760
5	0,01650	0,01840	0,02340	0,03000	0,03430	0,03830	0,04750	0,06180	0,07620	0,08530
4	0,01760	0,02030	0,02470	0,03040	0,03400	0,03670	0,05090	0,05840	0,07500	0,08790
3	0,01880	0,02080	0,02410	0,03090	0,03450	0,03700	0,04970	0,05490	0,08310	0,09340
2	0,01980	0,02270	0,02480	0,03350	0,03790	0,03650	0,04620	0,05070	0,08210	0,09370

Cuadro 17: Resultados Cllr Aitken-Lucy sin limitación

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00330	0,00260	0,00280	0,00240	0,00150	0,00210	0,00200	0,00340	0,00680	0,02180
9	0,00440	0,00400	0,00390	0,00350	0,00260	0,00300	0,00320	0,00340	0,00680	0,02180
8	0,00490	0,00420	0,00330	0,00300	0,00230	0,00310	0,00200	0,00340	0,00680	0,02180
7	0,00590	0,00520	0,00520	0,00520	0,00400	0,00420	0,00430	0,00340	0,00680	0,02180
6	0,00510	0,00440	0,00430	0,00340	0,00230	0,00310	0,00350	0,00340	0,00680	0,02180
5	0,00640	0,00600	0,00550	0,00460	0,00310	0,00350	0,00360	0,00340	0,00680	0,02180
4	0,00850	0,00810	0,00870	0,00840	0,00680	0,00690	0,00580	0,00340	0,00680	0,02180
3	0,01070	0,00930	0,00980	0,00980	0,00840	0,00990	0,00790	0,00340	0,00680	0,02180
2	0,01130	0,01630	0,01640	0,01540	0,01320	0,01450	0,01130	0,00850	0,01090	0,02160

Cuadro 18: Resultados Min_Cllr Aitken-Lucy sin limitación

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,01090	0,01190	0,01360	0,01820	0,02090	0,02540	0,03460	0,03740	0,06020	0,06280
9	0,01190	0,01380	0,01450	0,02020	0,02340	0,02810	0,03330	0,04090	0,05630	0,06580
8	0,01150	0,01350	0,01500	0,01920	0,02250	0,02950	0,03930	0,04300	0,05720	0,06760
7	0,01300	0,01350	0,01600	0,02140	0,02490	0,03030	0,03830	0,04560	0,05920	0,06990
6	0,01380	0,01430	0,01900	0,02410	0,02850	0,03080	0,04160	0,04830	0,06030	0,07730
5	0,01500	0,01600	0,02030	0,02560	0,02990	0,03390	0,04310	0,05510	0,06630	0,07600
4	0,01710	0,01790	0,02160	0,02600	0,02990	0,03230	0,04650	0,05170	0,06510	0,07760
3	0,01730	0,01840	0,02100	0,02650	0,03010	0,03260	0,04530	0,04820	0,07320	0,08310
2	0,01830	0,02030	0,02170	0,02910	0,03350	0,03210	0,04180	0,04400	0,07220	0,08340

Cuadro 19: Resultados Cllr Aitken-Lucy con limitación

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,00460	0,00330	0,00370	0,00410	0,00540	0,00530	0,00370	0,00340	0,00680	0,02180
9	0,00470	0,00400	0,00420	0,00450	0,00530	0,00530	0,00370	0,00340	0,00680	0,02180
8	0,00490	0,00420	0,00440	0,00470	0,00550	0,00550	0,00370	0,00340	0,00680	0,02180
7	0,00590	0,00510	0,00540	0,00550	0,00650	0,00640	0,00480	0,00340	0,00680	0,02180
6	0,00610	0,00540	0,00600	0,00610	0,00730	0,00730	0,00530	0,00600	0,00680	0,02180
5	0,00770	0,00690	0,00770	0,00760	0,00930	0,00820	0,00770	0,00800	0,00680	0,02180
4	0,00830	0,00710	0,00790	0,00760	0,00900	0,00910	0,00800	0,00620	0,00680	0,02180
3	0,01020	0,00900	0,00980	0,01020	0,01130	0,01270	0,01260	0,01340	0,00680	0,02180
2	0,01510	0,01410	0,01320	0,01330	0,01490	0,01680	0,01410	0,01380	0,01680	0,02180

Cuadro 20: Resultados Min_Cllr Aitken-Lucy con limitación

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,01380	0,01440	0,01540	0,01610	0,01720	0,01850	0,01810	0,01960	0,01700	0,04000
9	0,01940	0,01840	0,02040	0,01990	0,02070	0,03320	0,02310	0,01990	0,01870	0,03530
8	0,02680	0,02600	0,02930	0,02930	0,03300	0,03540	0,03140	0,02830	0,02460	0,04090
7	0,02490	0,02460	0,02530	0,02580	0,03130	0,03330	0,03220	0,02490	0,02610	0,05270
6	0,02150	0,02050	0,02190	0,02260	0,02480	0,02560	0,02590	0,02500	0,02590	0,04260
5	0,03510	0,03520	0,03870	0,03860	0,04300	0,03070	0,02710	0,02720	0,03130	0,05790
4	0,03450	0,03720	0,04090	0,04220	0,04750	0,04990	0,03010	0,02650	0,02880	0,05950
3	0,03240	0,02710	0,02910	0,02980	0,03310	0,03980	0,04050	0,04300	0,03110	0,05660
2	0,04010	0,04210	0,04070	0,04720	0,04070	0,04560	0,04570	0,04700	0,05140	0,06920

Cuadro 21: Resultados Cllr NFB

	Número de objetos									
M	200	180	160	140	120	100	80	60	40	20
10	0,01070	0,01100	0,01370	0,01370	0,01470	0,01680	0,01640	0,01660	0,01530	0,02150
9	0,01440	0,01330	0,01470	0,01380	0,01490	0,01820	0,01440	0,01440	0,00980	0,02180
8	0,02290	0,02150	0,02440	0,02330	0,02620	0,02810	0,02470	0,01030	0,00980	0,02180
7	0,02140	0,02080	0,02130	0,02130	0,02400	0,02730	0,02580	0,01590	0,01280	0,03330
6	0,01750	0,01610	0,01730	0,01720	0,01910	0,01770	0,01690	0,01850	0,01640	0,02180
5	0,02600	0,02520	0,02750	0,02680	0,02930	0,02470	0,01960	0,01760	0,02140	0,03720
4	0,02530	0,02300	0,02490	0,02430	0,02640	0,02670	0,02160	0,01500	0,01270	0,03330
3	0,02120	0,01910	0,02060	0,02090	0,02180	0,02610	0,02480	0,02230	0,02510	0,02180
2	0,02270	0,02150	0,02260	0,01920	0,02140	0,02160	0,01580	0,01400	0,02080	0,02160

Cuadro 22: Resultados Min_Cllr NFB